Computational analyses of transposable element target site preferences in *Drosophila melanogaster*.

A thesis submitted to The University of Manchester
for the Degree of PhD
in the Faculty of Life Sciences

2011

Raquel S. Linheiro

Table of Contents

(Word Count: 42,682)

List of Figures

List of Tables

Abstract

The University of Manchester

Raquel dos Santos Linheiro

Doctor of Science (DSc)

2011

Computational analyses of transposable element target site preferences in *Drosophila melanogaster*.

Transposable elements (TEs) are mobile DNA sequences that are a source of mutations and can target specific sites in host genome. Understanding the molecular mechanisms of TE target site preferences is a fundamental challenge in functional and evolutionary genomics. Here we used accurately mapped TE insertions in the *Drosophila melanogaster* genome, from large-scale gene disruption and resequencing projects, to better understand TE insertion site mechanisms. First we test predictions of the palindromic target site model for DNA transposon insertion using artificially generated P-element insertions. We provide evidence that the P-element targets a 14 bp palindromic motif that can be identified at the primary sequence level that differs significantly from random base composition in the *D. melanogaster* genome. This sequence also predicts local spacing, hotspots and strand orientation of P-element insertions. Next, we combine artificial P-element insertions with data from genome-wide studies on sequence properties of promoter regions, in an attempt to decode the genomic factors associated with P-element promoter targeting. Our results indicate that the P-element insertions are affected by nucleosome positioning and the presence of chromatin marks made by the *Polycomb* and *trithorax* protein groups. We provide the first genome-wide study which shows that core promoter architecture and chromatin structure impact P-element target preferences shedding light on the nuclear processes that influence its pattern of TE insertions across the *D. melanogaster* genome. In an effort to understand the natural insertion preferences of a wide range of TEs, we then used genome resequencing data to identify insertions sites not present in the reference strain. We found that both Illumina and 454 sequencing platforms showed consistent results in terms of target site duplication (TSD) and target site motif (TSM) discovery. We found that TSMs typically extend the TSD and are palindromic for both DNA and LTR elements with a variable center that depends on the length of the TSD. Additionally, we found that TEs from the same subclass present similar TSDs and TSMs. Finally, by correlating results on P-element insertion sites from natural strains with gene disruption experiments, we show that there is an overlap in target site preferences between artificial and natural insertion events and that P-element targeting of promoter regions of genes is a natural characteristic of this element that is influenced by the same features has the artificially generated insertions. Together, the results presented in this thesis provide important new findings about the target preferences of TEs in one of the best-studied and most important model organisms, and provide a platform for understanding target site preferences of TEs in other species using genomic data.

Declaration

No part of this thesis has been submitted in support of an application for any degree or qualification of The University of Manchester or any other University or Institute of learning.

Submission in Alternative Format

This thesis has been submitted in alternative format with permission from the Faculty of Life Sciences Graduate Office.

Copyright Statement

Acknowledgements

The first person I meet during my PhD was my supervisor Casey Bergman, who created the conditions for me to embark on this journey. He has built the foundations for my scientific input on both the bioinformatics and biological fields, for him a big thank you. I am also grateful to past Bergman lab members and specially those that are currently part of the group Maximilian Haussler, Pedro Olivares-Chauvet and Martin Gerner. Two previous members of the group stand out Ian Donaldson for being present from day one till now and Dave Gerrard for all the many conversations in these last years.

For all the academic, technical support and for managing the servers I wish to thank Nick Gresham and Adam Huffman.

I also whish to thank Sam Griffiths-Jones, Stefan Roberts, Matthew Ronshaugen, Ed Ryder and Steve Russell for stimulating discussion throughout the project and Don Rio and Roger Hoskins for helpful comments on the Chapter 2. I would also like to thank the Fundação para a Ciência e Tecnologia (FCT) from Portugal for funding my work.

I could not be here without the support of my family in particular my mother Luísa Linheiro, my grandparents and Suzete, I wish to give them a warm full big hug. Finally I would like to thank John Archer for being my family here, making me feel at home, putting up with my many moods and for all the endless theoretical and non-theoretical discussions.

Thanks,

Raquel

List of Abbreviations

bp	base pair
CAGE	cap analysis of gene expression
ChIP	Chromatin ImmunoPrecipitation
DCPD	*Drosophila* Core Promoter Database
DGDP	*Drosophila* Genome Disruption Project
DGRP	*Drosophila* Genetic Reference Panel
DPE	Downstream Promoter Element
DRE	DNA Replication Element
DSP1	Dorsal switch protein
FRT	Flip Recombinase Target
GAF	GAGA Associated Factor
GTF	General Transcription Factors
H3K27me3	trimethylation of Histone H3 on lysine 27
H3K4me3	trimethylation of Histone H3 on lysine 4
Hsp	Heat Shock Protein
Inr	Initiator
IR	Inverted Repeat
IR BP	Inverted Repeat Binding Protein
Kb	Kilobase pairs
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
modENCODE	model organism Encyclopedia of DNA elements
MTE	Motif Ten Element
MYA	million years ago
NHEJ	Non-Homologous End Joining
NURF	NUcleosome-Remodeling Factor
PC	Polycomb
PcG	Polycomb Group
PFM	Position Frequency Matrix
PH	Polyhomeotic
PHO	Pleohomeotic
PHOL	PHO like
PWM	Position Weight Matrix
RLM-RACE	RNA ligase mediated rapid amplification of cDNA ends
SDSA	Synthesis-Dependent Strand Annealing
SINE	Short Interspersed Nuclear Element
TBP	TATA-box Binding Protein
TE	Transposable Element
TIR	Terminal Inverted Repeats
TRF2	TATA-box Replication Factor 2
trxG	Trithorax Group
TSD	Target Site Duplication
TSM	Target Site Motif
TSS	Transcription Start Site
UTR	Untranslated Region

Thesis Outline

This thesis consists of three sections: a general introduction (Chapter 1), the body of the thesis comprising three distinct projects on the computational analysis of transposable element targeting (Chapters 2-4) and a conclusion section containing a summary of the main findings of this thesis and a discussion of future areas of research (Chapter 5).

Chapter 1 presents an overview of the major themes that are addressed in this thesis and it is divided in to five major subjects. Initially, a brief introduction to transposable elements (TE) is provided that primarily focuses on the *Drosophila melanogaster* P-element and its general features, usage and the current knowledge about its target site preference. This chapter also introduces a review of core promoters in *D. melanogaster* including their structure, motif composition and computational analysis. In the last two, sections a general overview of next generation sequencing is presented, along with a brief overview of some recurrent bioinformatics tools that will be used in the subsequent chapters.

Chapter 2 presents a minimally-revised version of the published work in Linheiro and Bergman (2008) "Testing the palindromic target site model for DNA transposon insertion using the *D. melanogaster* P-element." *Nucleic Acids Research* 36(19): 6199-208. In this chapter, we create a resource of reliably mapped artificial P-element insertions that is used as basis for all subsequent analysis in Chapters 2 and 3. This chapter also refines and characterizes a target site motif (TSM) identified at the sequence level that helps to clarify some aspects of P-element insertion such as local spacing, hotspots and strand orientation. Additionally, we show how this motif can be used to assess the quality of genome mappings for different P-element insertion libraries.

Chapter 3 presents unpublished work on the promoter target preferences of the P-element and focuses on the remarkable preference that the P-element demonstrates for the promoter region of genes. By combining previously selected P-element insertions with data from genome-wide studies on sequence properties of promoter regions, we attempt to decode the genomic factors associated with P-element promoter targeting. This analysis revealed a strong correlation of P-element targeting with some known

promoter motifs and clarifies the fine-scale spatial aspects of P-element promoter targeting. Additionally this chapter also reveals associations of P-element insertion with the presence of chromatin marks and tissue specific gene expression.

Chapter 4 presents unpublished work on the natural target site preferences of a large number of *D. melanogaster* TE families, using genome resequencing data of natural strains collected by the *Drosophila* Genetic Reference Panel. We show that next-generation sequencing reads from both Illumina and 454 sequencing platforms, can be used to discover target site duplications (TSD) and TSMs. Analysis of the TSD and TSM features from these different TE families revealed common properties of TE target site preferences for different classes of TE.

Chapter 5 presents a brief review of the major findings discussed in the 3 previous chapters, and also attempts to integrate results from chapter 3 and 4 in an effort to further clarify P-element promoter targeting preferences in natural strains.

1 Introduction

This chapter provides a general overview of the different topics in genetics and bioinformatics that are going to be covered in this thesis. Sections 1.1 and 1.2 provide an introduction to TEs and their methods of transposition, with a special emphasis on the *D. melanogaster* P-element and its role in *Drosophila* genetics. In these two sections, we discuss the major aspects that are currently known about P-element transposition. Section 1.3 discusses current knowledge about *D. melanogaster* core promoters, which are one of the major targets of P-element insertion. Section 1.4 provides a brief overview of the 454, Illumina and SOLiD next generation sequencing technologies, which are emerging as major sources of data to study TE insertion at the genome scale. Finally, section 1.5 provides an introduction to the major types of bioinformatics tools that were widely used in the data analysis for this thesis.

1.1 Transposable elements

TEs are mobile DNA elements that inhabit a host genome and whose movement can cause mutations and influence genome dynamics. They are considered a source of genetic variability and can be found in almost every organism, from prokaryotes to eukaryotes (Biemont and Vieira 2006). TEs can be divided into two major classes according to their method of transposition: (i) those that transpose through an RNA intermediate (retrotransposons) and (ii) those that transpose directly in to the host genome *via* a DNA intermediate (transposons) (Craig 2002; Biemont and Vieira 2006).

1.1.1 RNA-based transposition

Retrotransposons are characterized by transposition through an RNA intermediate and autonomous retrotransposons encode a reverse transcriptase (Biemont and Vieira 2006). Reverse transcriptase, which transcribes RNA into a DNA molecule that will integrate into the host genome, also works as an endonuclease, nicking the host DNA molecule at the target site (Craig 2002; Biemont and Vieira 2006). Retrotransposons can be divided into two major subclasses, those that have long terminal repeats (LTR) and those that do not (Craig 2002; Biemont and Vieira 2006). The non-LTR elements can be further divided into long interspersed nuclear elements (LINE) and short interspersed nuclear elements (SINE) (Craig 2002; Biemont and Vieira 2006). There are over 60 LTR and 40 non-LTR retrotransposon families currently documented in the *D. melanogaster* genome (Kaminker, Bergman et al. 2002). The *D. melanogaster* genome presents one of the most diverse organisms in terms of non-LTR element subclasses (Kapitonov and Jurka 2003). LTR elements are the most abundant class, comprising 2.65% of the euchromatin with non-LTR elements only occupying 0.87% of the euchromatin (Kaminker, Bergman et al. 2002).

1.1.2 DNA-based transposition

Transposons are bound by terminal inverted repeats (TIR) of variable length, which is one of their main defining features (Biemont and Vieira 2006). Full-length autonomous transposons, referred to by McClintock as activators, have in their coding sequence all the necessary information for the production of the proteins involved in transposition (Craig 2002). Non-autonomous transposons referred to by McClintock as dissociaters

are able to excise and transpose into the host genome, but lack the ability to produce its transposase protein, which is supplied by autonomous elements (Craig 2002).

In transposons, the most common transposition method is a "cut-and-paste" system that can be divided into two main events – excision from the donor site and insertion into the host genome. The mechanisms of excision are better understood than those of insertion. In eukaryotes one of the best-studied excision pathways is that of the *D. melanogaster* P-element (see Section 1.2).

Target selectivity is an important part of the transposition pathway and ultimately determines the type of mutation or rearrangement the insertion will cause. Insertions are typically thought to be non-random, since random integration events would ultimately lead to genome disruption and undermine transposon proliferation (Deininger and Roy-Engel 2002). Target selectivity varies among transposons and can be very specific as it is for the Tn7 element, which inserts itself into a specific site in the *Escherichia coli* genome (Bainton, Gamas et al. 1991). Conversely, target site selection may be less discriminative, as for the vertebrate Sleeping Beauty and insect Minos elements, which prefer AT-rich regions (Vigdal, Kaufman et al. 2002; Metaxakis, Oehler et al. 2005). One characteristic that seems to be prevalent among some transposons is a preference for palindromic or symmetrical sequences (see Chapter 4). A palindromic target sequence is consistent with most transposase proteins working in the form of multimers, choosing the target site on both DNA strands simultaneously (Rio 2002).

1.2 The *Drosophila* P-element

D. melanogaster, commonly known as the fruit fly, is one of the most powerful organisms in genetic analysis. One major advantage in flies is that its genome has few gene duplications (Kornberg and Krasnow 2000) allowing a rapid identification of mutated genes. Although there is a large evolutionary gap between flies and humans around 61% of the genes that in humans present an alteration, mutation or duplication in certain diseases have an orthologue in flies (Rubin, Yandell et al. 2000). One of the most powerful approaches to creating and mapping new mutations in *D. melanogaster* involves the use of transposon mutagenesis, principally involving the P-element transposon (Ashburner, Golic et al. 2005).

The P-element is thought to have recently incorporated into the *D. melanogaster* genome through horizontal transfer from *Drosophila willistoni* most likely in the early 20th century (Figure 1.1) (Daniels, Peterson et al. 1990). P-elements have rapidly spread through natural populations and in a short time, the only strains that remained P-element free were those kept in laboratory stocks (Engels, Johnson-Schlitz et al. 1990). The existence of two populations, one with P-elements and another without, allowed for a series of molecular and genetic techniques to be developed (see below) that fuelled interest in P-element biology.

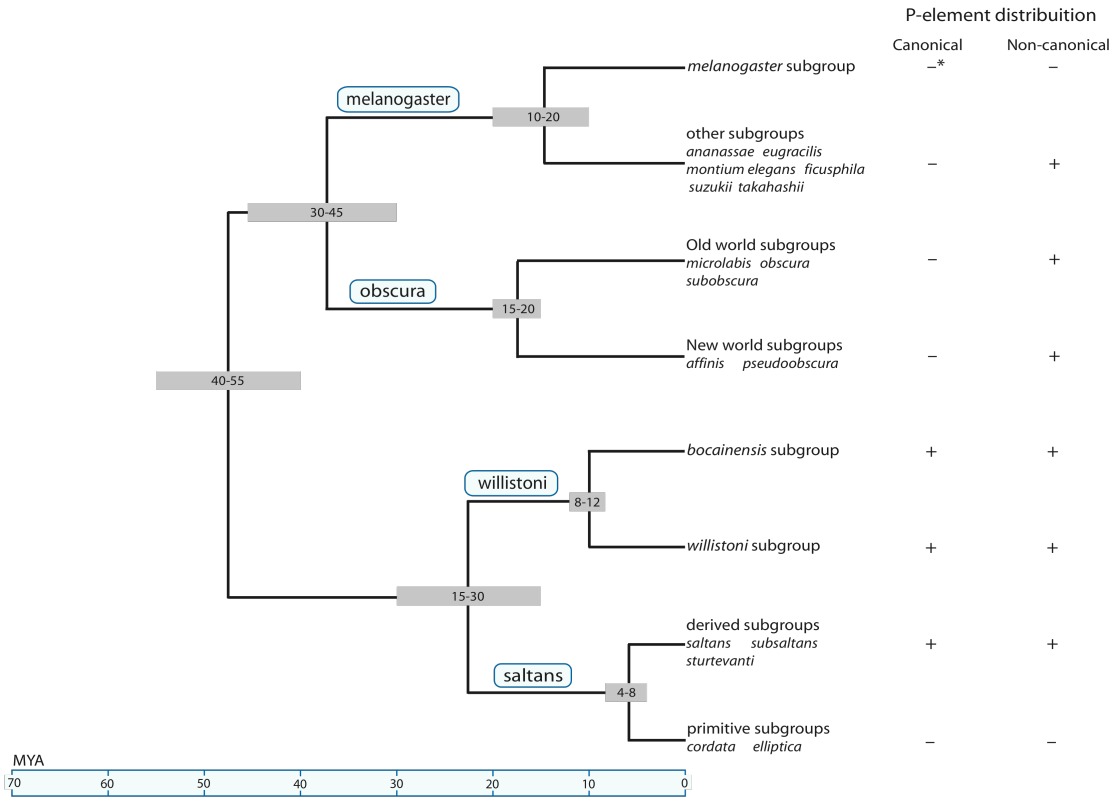


Figure 1.1 Phylogeny of the subgenus Sophophora

Phylogenetic tree adapted from (Clark and Kidwell 1997) showing the timeline of divergence and evolutionary relationships for the 4 main groups of the subgenus Sophophora alongside with the distribution of the canonical (complete and active elements similar to the one found in *D. melanogaster*) and the non-canonical P-elements. Canonical P-elements are found in both the *saltans* and *willistoni* groups, absent in the *obscura* group and only present in *D. melanogaster* from the *melanogaster* group suggesting horizontal transfer from the *willistoni* group into *D. melanogaster*.

Research into P-element biology led to the discovery of two cytoplasmatic types (cytotypes) that can be linked with the presence or absence of active P-elements in embryos. When P-elements are quiescent they are said to be in the P cytotype, which represents the cellular environment of eggs produced by mothers containing the P-element. When they are able to transpose they are said to be in the M cytotype, a state found in eggs produced by mothers that do not contain the P-element (reviewed in Karess and Rubin 1984).

When flies from a P and M cytotypes are crossed, a process known as hybrid dysgenesis occurs. This process only occurs when the P cytotype fly is a male (P strain or paternally contributing strain) and the M cytotype fly is a female (M strain or maternally contributing strain). When a P strain female is crossed with an M strain male, or when two P strains or two M strains are crossed, the hybrid dysgenesis process

does not take place (Rubin, Kidwell et al. 1982). Hybrid dysgenesis is confined to germline, where it induces a series of symptoms such as high rates of sterility, male recombination and chromosomal rearrangements (Spradling and Rubin 1982). The various phenotypes of hybrid dysgenesis are thought to result from the mobilization of P-elements in the fly genome causing double strand breaks in the genome.

In P strain flies, there are about 30 to 50 P-elements per haploid genome (O'Hare and Rubin 1983). One third of the elements are full-length P-elements that are able to encode the transposase protein. The remaining copies are truncated P-elements (0.5 to 1.6 Kb long) derived from the complete P-elements (O'Hare and Rubin 1983), that are non-autonomous elements some of which encode truncated version of transposase that acts as repressors of transposition (see section 1.2.4).

1.2.1 P-element structure

The complete *D. melanogaster* P-element, shown in Figure 1.2, is a DNA sequence of 2907 nucleotides long bound by perfect TIRs 31 base pairs (bp) wide and internal inverted repeat (IR) of 11 bp (O'Hare and Rubin 1983; Mullins, Rio et al. 1989).

The coding capacity of the full-length P-element resides in four large exons (numbered 0-3 for historical reasons), transcribed by a single transcript. According to different alternative splicing patterns, this transcript may give rise to the transposase protein, if it is completely translated, or the truncated 66kDa repressor protein, when exon 3 is not translated (O'Hare and Rubin 1983; Karess and Rubin 1984; Rio, Laski et al. 1986; Misra and Rio 1990).

The transcription of the P-element transposase is under the control of a TATA-box motif containing RNA polymerase II promoter. The P-element TATA box overlaps the P-element transposase binding site (see below) and has the consensus sequence – TTAAATT (Kaufman, Doll et al. 1989). This is a weak promoter that can be put under the control of other promoters, a useful attribute that has been exploited in several techniques (see Section 1.2.6) (Golic and Lindquist 1989; Brand and Perrimon 1993; Ward, Thaipisuttikul et al. 2002).

The P-element also contains two transposase binding sites (TBS), which are present at the 5' and 3' ends of the element in inverted orientation, adjacent to the 31 bp TIRs overlapping the P-element TATA box (Figure 1.2) (Kaufman, Doll et al. 1989). These 10 bp consensus sequences (ATMCACTTAA) are positioned unevenly in the P-element, 21 bp apart from the 31 bp TIR at the 5' end and 9 bp from the 3' end (Kaufman, Doll et al. 1989; Beall and Rio 1997).

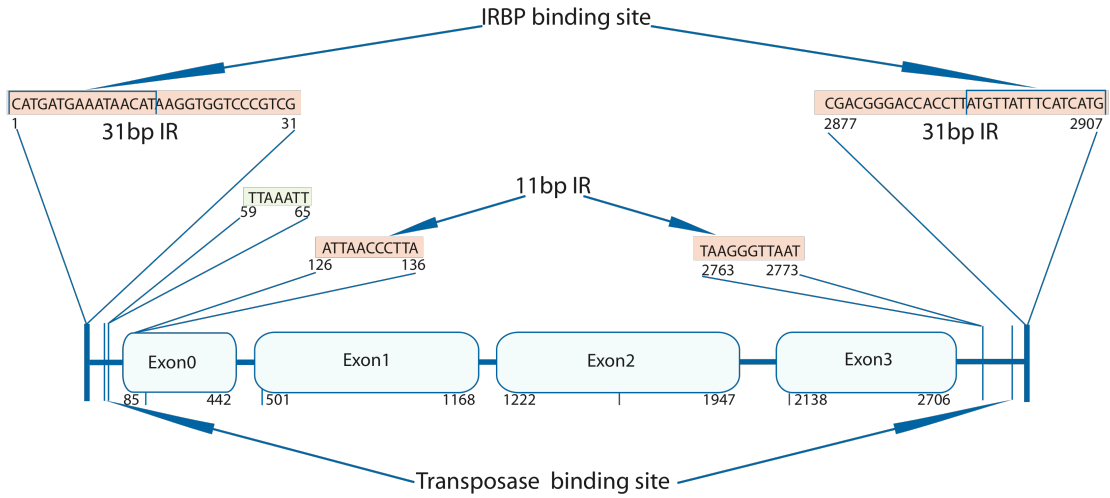


Figure 1.2 Structure of a complete P-element

A full length P-element is 2907 bp long and contains a single transcription unit with four exons (rounded boxes). The numbers under the exons indicate their start and end positions in the full-length element. The sequences of the 31 bp inverted repeats flanking the P-element ends are highlighted, as are the location of the 11 bp inverted repeats inside the body of the P-element. Sites of known protein binding are also shown, including the 10 bp transposase binding sites and the 16 bp inverted repeat binding protein (IR BP) binding sites, located at the 5' and 3' end of the P-element, adapted from (Rio 2002). The weak TATA-box motif is also indicated at the top of the element in light green (Kaufman, Doll et al. 1989).

1.2.2 P-element transposition

The P-element transposition pathway is through a cut-and-paste system that requires the transposase protein and essential DNA sequences in the P-element 5' and 3' terminal regions. These include the TIR located at both P-element ends, the transposase binding site (TBS) and the inverted repeat binding protein (IR BP) sites (Kaufman, Doll et al. 1989; Mullins, Rio et al. 1989). Structural differences in the P-element 5' and 3' ends make them non-interchangeable, and a P-element requires both 5' and 3' ends for a normal transposition event (Mullins, Rio et al. 1989). Outside the essential terminal

regions there is the 11 bp wide inverted repeat (IR) that functions as an enhancer of the transposition process (Figure 1.2) (Mullins, Rio et al. 1989).

The P-element transposase protein is an 87 kDa protein containing six functional regions (Figure 1.3): the DNA binding domain (a C_2HC putative zinc motif contiguous with a basic region), a potential phosphorylation site, two dimerization regions, a Guanosine Triphosphate (GTP) binding motif and a catalytic region (Kaufman, Doll et al. 1989; Rio 2002). The catalytic region has a non-canonical motif that has four acidic residues (DDED) instead of three acidic residues of a typical transposase DDE motif.

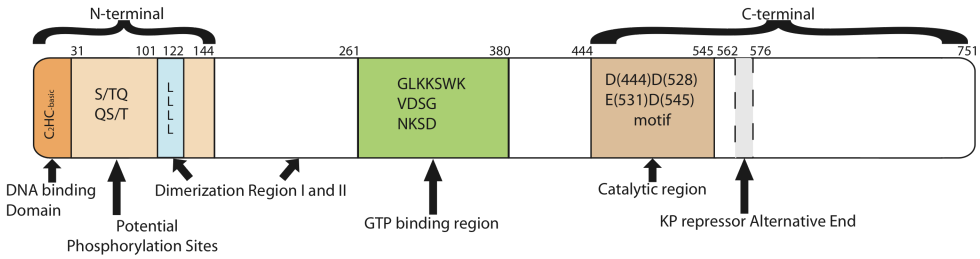


Figure 1.3 Structure of P-element transposase protein

From left to right, the DNA binding domain at the N-terminus; potential phosphorylation sites for the ATM family of DNA repair-checkpoint phosphatidyl inositol-3-phosphate related proteins; two dimerization regions: first a leucine zipper, and second a domain with no known motif; GTP binding sites with the most conserved motifs in the GTPase superfamily; and the catalytic region in the C-terminal with an atypical DDED motif. The numbers on top indicate the amino acid position of the domains (Rio 2002).

The P-element transposition process starts when the P-element transposase protein binds to the TBS and associates with the 31 bp TIR, bringing together the P-element ends at the donor site (Beall and Rio 1997). Transposase then creates a 17-nucleotide wide staggered cut that exposes the reactive 5' and 3' ends of the P-element (Figure 1.4) (Beall and Rio 1998). The excised P-element together with transposase protein then forms a stable hairpin shaped complex that will interact with the target site (Tang, Cecconi et al. 2005).

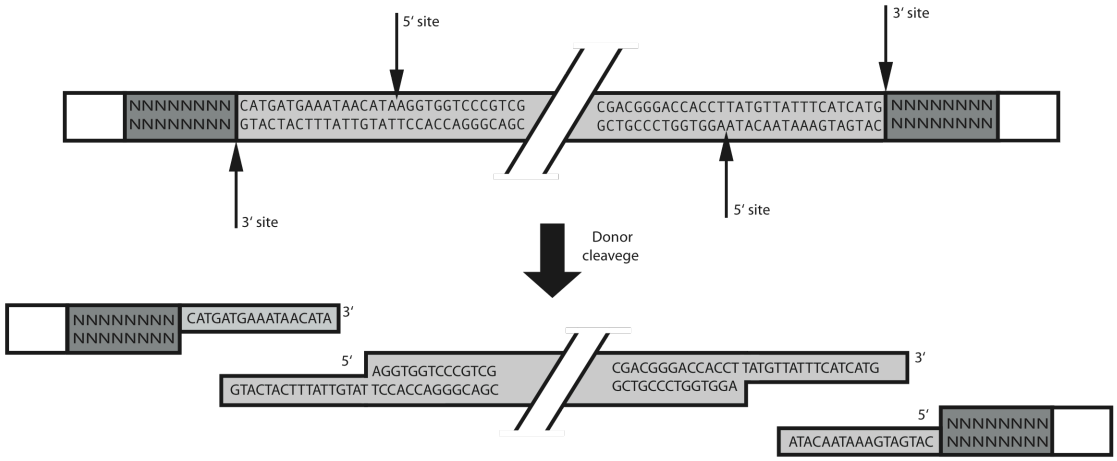


Figure 1.4 P-element staggered cuts at the donor site

Top diagram represents the P-element at the donor site with the 31 bp TIR (light grey), the TSDs at both ends (dark grey) and the transposase cut sites which are indicated by the 4 arrows. The bottom diagram shows the excised P-element and the donor site with the exposed staggered cuts with 3' overhangs. The reactive P-element ends will interact with a new target site and the exposed donor sites will be repaired either by non-homologous end joining (NHEJ) or by synthesis-dependent strand annealing (SDSA). Diagram adapted from (Beall and Rio 1997).

1.2.3 Cofactors involved in P-element transposition

There are several factors involved in the P-element transposition in addition to the transposase protein and sequences in the P-element. Proteins such as the IR BP, topoisomerases and histone-like proteins may also be involved in P-element transposition (Mullins, Rio et al. 1989). Other important factors to the P-element transposase are the nucleotide Guanosine-5'-triphosphate (GTP) and the metal magnesium ions (Mg^{2+}) (Kaufman and Rio 1992).

GTP is an essential nucleotide in P-element transposition. Several experiments (Kaufman and Rio 1992; Mul and Rio 1997; Tang, Cecconi et al. 2005) have shown that GTP acts as a regulatory molecule during the formation of the synaptic complex between transposase and P-element. The requirement of GTP makes it unique among the family of transposase/integrase proteins, and its concentration levels may help to regulate the transposition reaction (reviewed Mul and Rio 1997; Tang, Cecconi et al. 2005).

Although GTP is essential to transposition, its hydrolysis is not required. Experiments have shown that Guanosine-5'-diphosphate (GDP) could also assist in the transposition reaction (Kaufman and Rio 1992). The non-requirement of energy during transposition is supportive of the cut-and-paste transposition mechanism (Kaufman and Rio 1992; Mul and Rio 1997). The GTP binding motif in transposase is unique among the GTP family and mutations in this region prevent transposition (Mul and Rio 1997).

1.2.4 Repressors of P-element transposition

P-element repression in P cytotype flies is mainly associated with repressor proteins encoded by truncated versions of the P-element transposase gene. Two main repressor proteins are the 66 kDa and the KP repressor proteins. The 66-kDa repressor protein is believed to be one of the components of the P cytotype in flies and can be found in the oocytes of P females (reviewed in Rio 2002). This protein results from alternative splicing of the full-length P-element transcript that lacks ORF3 (reviewed in Rio 2002). The KP repressor protein is a 207 aa version of the transposase protein encoded by a naturally occurring truncated P-element (Figure 1.3) (Lee, Mul et al. 1996). Along with the 66-kDa repressor and transposase proteins, the KP repressor protein recognizes the P-element binding sites in the 11 and 31 bp IRs (Lee, Mul et al. 1996; Rio 2002). The first 88 aa of the KP repressor binds to the transposase binding site and contains a CCHC motif which is a putative metal binding site (Lee, Beall et al. 1998). This motif is shared by the transposase and the 66-kDa repressor proteins and is essential to site specific DNA binding (Lee, Beall et al. 1998). As with the transposase protein, the KP repressor also possesses a leucine zipper that is responsible for its dimerization (Lee, Mul et al. 1996) but is not essential for its inhibitory effects (Lee, Beall et al. 1998).

It is believed that inhibition of transposition by the repressor proteins occurs by 3 processes – impairment of the transcription; binding to the TIR of the P-element, and by binding to the transposase binding site (reviewed in Rio 2002). The existence of these proteins or similar ones does not entirely explain the mechanism of repression, which is still not fully understood (Rio 2002).

1.2.5 P-element gap repair mechanism

During P-element transposition, it is necessary to repair both the donor site and the target site. At the target site the host repair mechanism generates an 8 bp TSD adjacent

to the 31 bp TIR (O'Hare and Rubin 1983; Kaufman, Doll et al. 1989; Engels, Johnson-Schlitz et al. 1990). There are two main processes by which the gap left at the donor site can be repaired: non-homologous end joining (NHEJ) and synthesis-dependent strand annealing (SDSA) (Rio 2002). Each of these involves IR BPs, such as the Ku P70 and Ku P80 proteins (reviewed in Rio 2002), which stabilize the double strand break at the donor site (Staveley, Heslip et al. 1995).

NHEJ is a simple mechanism in which the two ends left at the donor site are joined. This process is accomplished by the IR BPs that are bound to the 17 nt staggered cut at the donor site (Staveley, Heslip et al. 1995; Rio 2002). In this mechanism there is no need for an extensive homology between the broken ends or an intact homologus chromosome (Smith and Jackson 1999). This process is associated with some of the precise excision events observed (Rubin, Kidwell et al. 1982; O'Hare and Rubin 1983; Engels, Johnson-Schlitz et al. 1990). Since NHEJ binds both 17 nt staggered cuts together, repair of the donor site will result in a few extra nucleotides, including the 17 nt from the staggered cut and the 8 bp TSD that were absent before P-element insertion.

In SDSA the two cleaved DNA ends search independently for a homologous sequence that will serve as a template for DNA synthesis (Nassif, Penney et al. 1994), often from the homologous chromosome or sister chromatid depending on when the transposition event happens. SDSA is linked with the formation of naturally truncated P-elements and with an increase in TE number when a TE is used as a template (Gloor, Nassif et al. 1991; Nassif, Penney et al. 1994).

1.2.6 P-element based genetic engineering applications

Understanding the P-element transposition and gap repair mechanisms has facilitated the development of a set of powerful genetic and genomic analysis tools in *Drosophila*. Examples include the production of stable lines with regulated transgenes, efficient production of genetic mosaics, screens for mutants with specific tissue and cell phenotypes, and observation of gene expression patterns (Kornberg and Krasnow 2000).

The most basic genetic engineering technique using the P-element is to generate insertion mutations. TE insertion mutations are often preferred over other types of mutagenic agents, since they allow rapid mutant cloning *via* plasmid rescue or inverse

PCR, mutation detection through specific phenotypes associated with marker genes in the transposon, and analysis of revertant mutations (reviewed by Spradling, Stern et al. 1995). These convenient features often make P-element mutagenesis a more desirable technique over other mutagenic agents such as radiation or chemicals.

P-element genetic techniques take advantage of several aspects linked with knowledge or P-element transposition. The P-element transposase source can be supplied in *trans* by a helper P-element – a P-element that is able to express the transposase protein but is unable to transpose because it lacks the TIRs (Karess and Rubin 1984). This allows the production of stable P-element mutations and transgenes that contain terminal regions but lack a transposase gene of their own. P-element mobilization can also be induced by transforming engineered P-element constructs into embryos and co-injecting transposase on helper plasmids (Spradling and Rubin 1982). The ability to introduce and mobilize genetically engineered P-element constructs allowed the development of techniques that combine the P-element with yeast flip recombinase protein (FLP) to produce recombination events (Golic and Lindquist 1989; Golic and Golic 1996; Rong and Golic 2000), with the *E. coli lacZ* gene for expression patterns analysis (Bellen, O'Kane et al. 1989) and the yeast GAL4-UAS system to misexpress genes (Golic and Lindquist 1989; Brand and Perrimon 1993; Ward, Thaipisuttikul et al. 2002).

Recombination events with the FLP are produced through the use of two P-element constructs. The first P-element encodes the FLP that is put under the control of the *hsp70* promoter (Golic and Lindquist 1989). The other P-element has a Flip recombinase target (FRT) and a marker gene such as the *white* gene that enables a phenotypic visualization of the mutations (Golic and Lindquist 1989). This technique allows cell lineage analysis (Golic and Lindquist 1989) and, by combining two P-elements with FRT sites in the proper orientation, to create large-scale chromosome rearrangements such as paracentric and pericentric inversions, duplications and deficiencies (Golic and Golic 1996).

Inclusion of a *cis*-regulatory element in a *lacZ* containing P-element allows reporter gene analysis (Hiromi, Kuroiwa et al. 1985). Mobilization of a *lacZ* containing P-element lacking any *cis*-regulatory element besides its promoter allows enhancer trap analysis (O'Kane and Gehring 1987). Since the P-element promoter is weak, it is

influenced by its surroundings and the *lacZ* expression patterns mimic its neighbouring genes (Bellen, O'Kane et al. 1989). This technique allows the rapid identification of expression patterns across developmental stages for genes that have a P-element insertion (Bellen, O'Kane et al. 1989).

The yeast GAL4-UAS system can be incorporated into P-elements to permit induced gene activation (Brand and Perrimon 1993; Ward, Thaipisuttikul et al. 2002). This technique uses two P-element constructs, one that carries the yeast GAL4 transcriptional activator that is put under the control of a promoter or enhancer sequence by an enhancer trap or reporter construct, and a second P-element elsewhere in the genome that carries GAL4 binding sites from yeast called upstream activator sequences (UAS) and a gene of interest or marker gene, such as *lacZ* (Brand and Perrimon 1993). When comparing to heat shock activation, the GAL4-UAS system provides more specific and controllable inducible mis-expression (Brand and Perrimon 1993).

These techniques have allowed the rapid identification and characterization of individual genes, expression patterns and mutations in *Drosophila*, and have motivated the construction of large-scale P-element insertion libraries for use in functional genomics. One such project was the *Drosophila* Gene Disruption Project (DGDP) that gathered a set of more than 7,100 strains of P-element and *piggyBac* insertions that were associated with 40% of *Drosophila* annotated genes (Bellen, Levis et al. 2004). Similarly, the DrosDel Collection generated a set of 3,242 P-element insertions for use in creating deletions (Ryder, Blows et al. 2004). Finally, a set of 3,825 insertions of the P{GawB} element has been generated for GAL4 enhancer trap analysis (Hayashi, Ito et al. 2002).

1.2.7 P-element target site preferences

The target site preferences of the P-element are complex and have been the subject of several studies (O'Hare and Rubin 1983; Roiha, Rubin et al. 1988; Bownes 1990; Berg and Spradling 1991; Tower, Karpen et al. 1993; Zhang and Spradling 1993; Liao, Rehm et al. 2000; Timakov, Liu et al. 2002; Julian 2003; Geurts, Hackett et al. 2006). From these studies the major aspects that have been inferred are a preference for insertion into the 5' UTR and promoter region of genes (Roiha, Rubin et al. 1988; Bellen, O'Kane et al. 1989; Tower, Karpen et al. 1993; Zhang and Spradling 1993; Timakov, Liu et al.

2002), euchromatic regions (Bellen, O'Kane et al. 1989) and genes with an open chromatin configuration like *heat shock protein* (*Hsp*) genes (Lerman, Michalak et al. 2003; Walser, Chen et al. 2006). Additionally, some general structural aspects of P-element insertion preference have also been inferred from the analysis of insertion site sequences (O'Hare and Rubin 1983; Liao, Rehm et al. 2000; Julian 2003).

The original study on P-element target sites conducted by (O'Hare and Rubin 1983) used a set of 18 insertions to demonstrate a preference for the target site (GGCCAGAC); however given the small size of the sample no evidence for a consensus sequence was attained. A study by (Liao, Rehm et al. 2000) based on 1,185 insertions sites found a 14 bp palindromic hydrogen-bonding pattern that appeared to be related to structural features of DNA, but concluded that "although there are base preferences at each position, these are not strong enough to generate a clear consensus sequence". A more recent research by (Julian 2003), with a dataset of 795 insertions sites, found a palindromic motif at the target site but reported a non-palindromic sequence (ANNGGCCAGACNNT) that agreed with the sequence found by (O'Hare and Rubin 1983).

In addition to the well-known tendency to insert in promoter regions, the P-element has been shown to insert near genes that are active in the germline (Bownes 1990; Timakov, Liu et al. 2002) and promoters of *Hsp* genes (Lerman, Michalak et al. 2003; Walser, Chen et al. 2006). The preference for *Hsp* promoters has been associated with the proposed open chromatin structure of these genes. Because *Hsp* genes have to act in a short period of time their promoters are thought to be "poised" for transcription, with decondensed chromatin and polymerase ready to elongate (Walser, Chen et al. 2006).

These known aspects of P-element target site selection are summarized in Table 1.1. Despite clear tendencies for non-random integration, the full set of factors that influence P-element target selection currently remains unknown.

Table 1.1 P-element target site preferences

Specific sequences	References
GGCCAGAC	(O'Hare and Rubin 1983; Roiha, Rubin et al. 1988; Preston, Sved et al. 1996; Beall and Rio 1998)
GTCCGGAC	(Liao, Rehm et al. 2000)
ANNGTCCGGACNNT	(Julian 2003)
Genomic Regions	
Euchromatic regions	(Bellen, O'Kane et al. 1989)
5' end of the original P-element	(Bellen, O'Kane et al. 1989; Tower, Karpen et al. 1993; Zhang and Spradling 1993; Timakov, Liu et al. 2002)
5' end of genes	(Roiha, Rubin et al. 1988; Zhang and Spradling 1993)
Promoter region of *hsp* genes	(Lerman, Michalak et al. 2003; Walser, Chen et al. 2006)
Structural Features	
DNA bendability	(Liao, Rehm et al. 2000)
A-philicity	(Liao, Rehm et al. 2000)
Protein induced deformability	(Liao, Rehm et al. 2000)
B-DNA twist	(Liao, Rehm et al. 2000)

1.3 Core Promoters

As noted in the previous section, one of dominant features of P-element target preferences is to insert into the promoter region of genes (Roiha, Rubin et al. 1988; Bellen, O'Kane et al. 1989; Tower, Karpen et al. 1993; Zhang and Spradling 1993; Timakov, Liu et al. 2002). Additionally, P-element insertions have been recovered in only a subset of around ~40% of *D. melanogaster* genes (Bellen, Levis et al. 2004). The mechanistic basis of both these aspects of P-element target preferences remain open questions, although they have been proposed to be related to core promoter architecture and function.

The core promoter is a nucleotide sequence usually located between 35 bp upstream and downstream of the Transcription Start Site (TSS). Its main function is to recruit the General Transcription Factors (GTFs) associated with RNA polymerase II, the polymerase responsible for messenger RNA (mRNA) transcription of protein coding genes.

A great diversity of core promoter types exists in *D. melanogaster*, some of which have been associated with different regulatory activities (Ohtsuki, Levine et al. 1998; Ohler 2006). This variation is likely related to the multiple processes that control gene expression in the wide range of tissues presented in higher eukaryotes (Ohler 2006). Different promoter types may be connected to different cellular pathways, especially those related with cellular development and differentiation (Down, Bergman et al. 2007), mechanisms of transcription, or 3-dimensional organization (Gershenzon, Trifonov et al. 2006).

1.3.1 Types of core promoters

Core promoters in eukaryotes differ widely in terms of their architecture, expression, and function. For example, two extreme types of promoters are those with focused or dispersed motifs. Focused promoters have characteristic motifs that are located in specific sites around the TSS; such motifs include the TATA box, the Initiator (Inr) sequence and the Downstream Promoter Element (DPE) (Juven-Gershon, Hsu et al. 2006). Dispersed promoters have a range of weak motifs that spread in a 50 to 150 nt

extension around the TSS and have been associated with TATA-less promoters and CpG islands (Juven-Gershon, Hsu et al. 2006).

Likewise, different types of promoters can be active in different developmental stages. TATA box containing promoters have been associated with post-embryonic development (Bajic, Tan et al. 2006; Muller, Demeny et al. 2007), while TATA-less promoters have been associated with early stages of development (Muller, Demeny et al. 2007). In mammals, promoters containing CpG islands were linked with housekeeping genes expressed in all types of tissues and cells (Aerts, Thijs et al. 2004).

Finally, different types of core promoters can have different functions. For instance the DPE motif has been associated with RNA polymerase pausing (Hendrix, Hong et al. 2008) a state characteristic of genes such as *Hsp*70 were the gene is transcribed in less then 60 seconds after heat shock (Gilmour 2009). Transcription factors that are part of the pre-initiation complex (see below) prefer certain types of promoters to others, and certain enhancers may present a preference for promoters containing specific promoter motifs, such as the TATA box (Ohtsuki, Levine et al. 1998).

1.3.2 Core promoter motifs

Many aspects of core promoter architecture are related to the sequence motifs that are involved in transcriptional initiation. Before RNA polymerase II can initiate transcription, a set of GTFs has to assemble at the TSS to form the pre-initiation complex. There are several known GTFs and those that are associated with RNA polymerase II are generally referred to as Transcription Factors for RNA polymerase II (TFII). For example, one of the main TFII GTFs is TFIID, which contains the TATA binding protein (TBP) that binds the TATA box motif (Juven-Gershon, Hsu et al. 2006; Muller, Demeny et al. 2007).

The TATA box is the best studied core promoter motif and is able to recruit the GTFs by itself (Juven-Gershon, Hsu et al. 2006). This motif is often found in an AT rich region from -20 to -30 bp upstream of the TSS (Kutach and Kadonaga 2000). Although considered one of the most common promoter motifs, it has only been associated with 15% of the human (Juven-Gershon, Hsu et al. 2006) and *D. melanogaster* core promoter regions (Gershenzon, Trifonov et al. 2006).

The most commonly observed promoter motif is the Inr element, which is found directly at the TSS. The Inr is unable to recruit the GTFs by itself and is often associated with other promoter elements such as the TATA box, DPE and Motif Ten Element (MTE) (Juven-Gershon, Hsu et al. 2006).

In *Drosophila* the DPE sequence appears to be as common as the TATA box motif (Kutach and Kadonaga 2000). This sequence is often found between 28 and 33 bp downstream of the TSS (Lim, Santoso et al. 2004) and is unable to recruit the GTFs by itself, but instead works in synergy with the Inr (Kutach and Kadonaga 2000). The DPE and Inr are strongly associated with each other in the genome and have a characteristic spacing between them (Kutach and Kadonaga 2000). When there is an alteration in the distance between DPE and Inr, binding of TFIID to the promoter is decreased and consequently the rate of transcription is also reduced (Kutach and Kadonaga 2000).

The MTE motif is similar to the DPE element (Lim, Santoso et al. 2004), and located in a similar position, ranging from 18 to 29 bp downstream of the TSS, overlapping the first base pairs of the DPE sequence (Lim, Santoso et al. 2004). As with the DPE, the MTE functions in synergy with the Inr sequence and the distance between these motifs (around 10 to 11 bp) is also an important factor in their capability to recruit the GTFs (Lim, Santoso et al. 2004). The MTE promoter can function either independently of the TATA box and DPE promoters, being able to substitute them, or in cooperation increasing their activity (Lim, Santoso et al. 2004).

1.3.3 Computational analyses of core promoter libraries in *Drosophila*

To shed light on the mechanism by which a promoter is regulated there has been substantial effort to use computational methods for finding new promoter regions and building promoter motif libraries. These studies have revealed many new promoter motifs and new aspects of core promoter biology. For example, one surprising finding was that the TATA box is a much less frequent promoter element than previously thought (Kutach and Kadonaga 2000; Gershenzon, Trifonov et al. 2006).

These studies have relied on collections of functionally characterized promoter regions, such as the *Drosophila* Promoter Database (DPD) with 247 TSS, and the *Drosophila* Core Promoter Database (DCPD) with 205 start sites (reviewed by Ohler 2006). Ohler

(2006) extended these collections by identifying 1,941 TSS based on 5' EST clusters, corresponding to 14% of all genes in *D. melanogaster*. Ohler's (2002) dataset only overlaps 18 and 16% of the DPD and DCDP collections (Ohler, Liao et al. 2002).

A more recent project, part of the model organism Encyclopedia of DNA elements (modENCODE) project, has collected a large set of TSSs from large-scale sequencing techniques such as cap analysis of gene expression (CAGE) and RNA ligase mediated rapid amplification of cDNA ends (RLM-RACE) (Hoskins, Landolin et al. 2010). This project generated TSS data for 12,454 promoter regions and characterized the shapes and TSS distribution of more then half of the annotated genes in *D. melanogaster.*

These TSS libraries have been essential for discovering key sequence features of promoters in *Drosophila*. For example, there is an increase in AT content just before the TSS in *Drosophila* that decreases to below average at the TSS (without changing AT predominance) (Aerts, Thijs et al. 2004). These data have also been used to show that promoters in *D. melanogaster* fall into two categories: those that initiate at a single peak (peaked) and those that form TSS clusters (broad) (Rach, Yuan et al. 2009; Hoskins, Landolin et al.). Promoter motifs differ in each of these categories of promoter with positionally flexible motifs like the DNA replication element (DRE) more frequent in broad promoters, and position-restricted motifs like TATA, Inr, DPE, pause button (PB) and GAGA more frequent in peaked promoters (Rach, Yuan et al. 2009; Hoskins, Landolin et al. 2010). Peaked promoters were also shown to be preferentially associated with developmentally restricted gene expression (Hoskins, Landolin et al. 2010).

1.4 High throughput DNA sequencing

Large-scale DNA sequencing has been an essential part of the in-depth analysis of all aspects of complex genomes (Myers, Sutton et al. 2000; Venter, Adams et al. 2001; Celniker, Wheeler et al. 2002; Warren, Hillier et al. 2008). DNA sequencing technologies have evolved substantially in the last few decades, from the time consuming and expensive Sanger dideoxy chain-termination method to high-throughput methods such as parallelized pyrosequencing and reverse termination techniques (Schuster 2008; Pettersson, Lundeberg et al. 2009). High throughput DNA sequencing technologies provide an explosion of data to better understanding genetic and epigenetic processes. A brief introduction to these technologies follows, to provide context for the high-throughput sequencing data used in this thesis.

1.4.1 Parallelized pyrosequencing

The first major high-throughput sequencing platform was the 454 Life Sciences Genome Sequencer from Roche that uses highly parallelized pyrosequencing to generate sequence outputs of over 1.25 million 400 bp reads in a single run (Pettersson, Lundeberg et al. 2009). It uses emulsion PCR for amplification and reaction between pyrophosphate molecules and the firefly luciferase enzyme with a sequential identification of the added base for detection (Figure 1.5 - Left panel) (Mardis 2008; Pettersson, Lundeberg et al. 2009). The first four bases in the adaptor (TCGA) calibrate the instrument for the processing of each base during the run (Mardis 2008). The pyrosequencing reaction and single nucleotide calibration process is not sensitive enough to allow for an exact number of bases in homopolymer runs errors (>6 of the same base) (Mardis 2008). Both the homopolymer errors and PCR-introduced errors at the sample level can be overcome with oversampling (>20 runs), but remain present in individual 454 reads (Pettersson, Lundeberg et al. 2009).

1.4.2 Reverse termination

Another recently developed high-throughput parallelized sequencing process is the Solexa/Illumina Genome Analyzer. The Illumina sequencing platform is based on bridge amplification and sequencing by synthesis with reversible chain termination with all four nucleotides added at the same time (Figure 1.5 – Right panel) (Pettersson, Lundeberg et al. 2009). The solid support for the Illumina sequencer is an eight-lane

flow cell that allows the analysis of eight different DNA libraries (Mardis 2008). The read length attained is dependent on the number of cycles allowed during sequencing which can now go up to 150 bp. As with the 454 sequencer a post-processing pipeline selects for poor quality reads that are automatically removed from the final output (Mardis 2008). In theory this method is similar to the Sanger method since it uses fluorescent reversible dye terminators, although the parallelized sequencing allows for a faster and lower cost sequencing (Medini, Serruto et al. 2008; Pettersson, Lundeberg et al. 2009).

1.4.3 Ligase mediated sequencing

In contrast with the two previous methods the Applied Biosystems SOLiD Sequencer is a polymerase independent method that relies on DNA ligase. Similarly though sequencing proceeds in an adapter-ligated fragment with PCR amplification of the sequences in small magnetic beads (Medini, Serruto et al. 2008;Mardis 2008). During the sequencing process 8 mer probes are added to a glass slide with all the 1024 possible combinations for the fist 5 bases (Janitz 2008). Bases 1 and 2 of the 8 mer are identified through a four-color dye code with the last 6 bases preventing further binding (Janitz 2008; Mardis 2008). At the end of each cycle the bound probes are chemically cleaved between positions 5 and 6 removing the last 3 bases and the fluorescent dyes (Janitz 2008). In the second and posterior cycles the next 2 bases are identified with a 3bp gap between identified dinucleotides, the process then continues for a number of user specified cycles (Janitz 2008; Mardis 2008). To identify the remaining bases in the sequence and to identify each single base independently the cycles are repeated with five different primers that go from *n* to *n*-4 considering the first primer *n* length (Janitz 2008; Mardis 2008). Since the dyes are associated with dinucleotides each base is defined twice by the color code, conferring a higher accuracy in the identification of each base (Janitz 2008; Mardis 2008). The major limiting factor of this technique seems to be the read length which is much smaller then in the two previous methods, varying from 25 to 35bp for 5 to 7 sequencing cycles respectively (Janitz 2008; Mardis 2008).

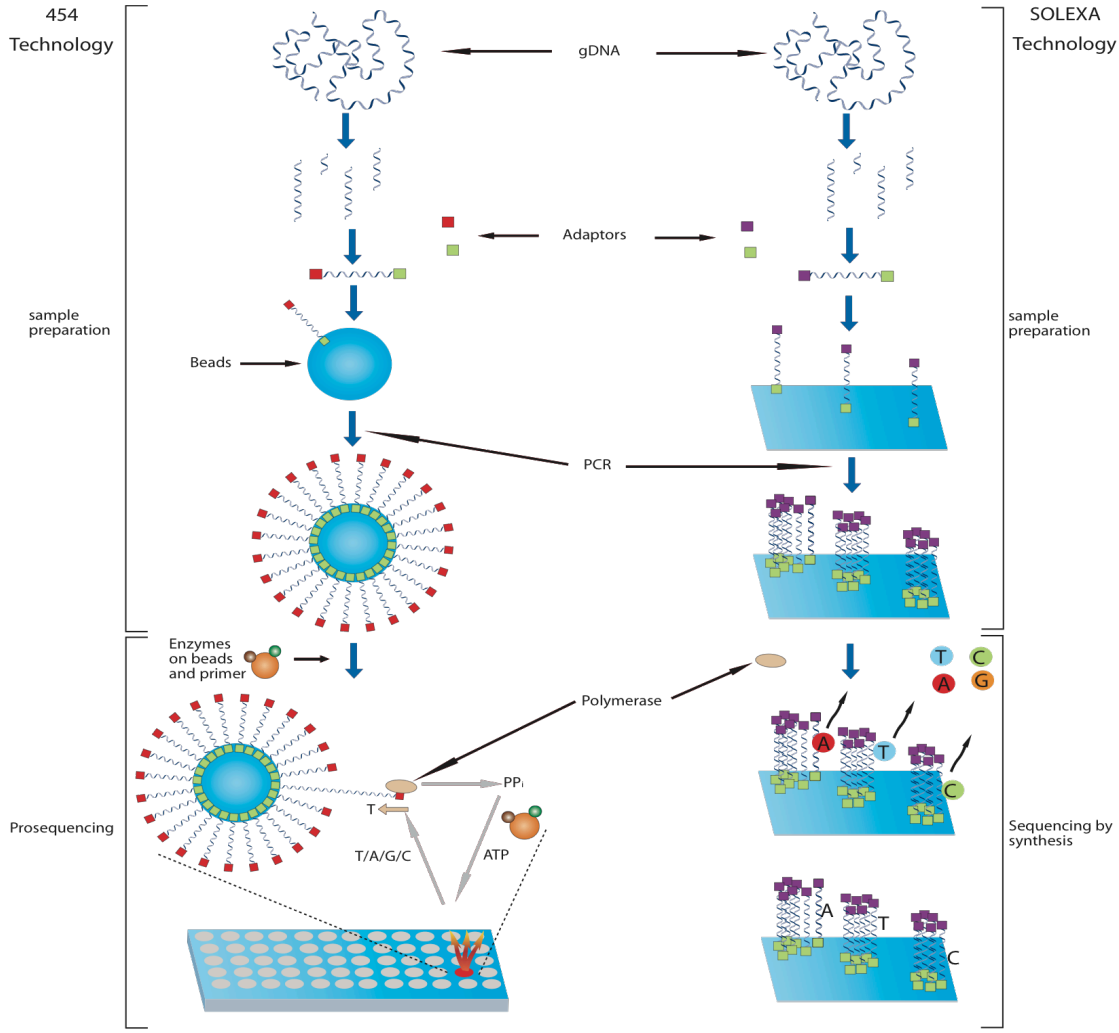


Figure 1.5 Schematics of parallelized 454/Roche pyrosequencing and Solexa/Illumina reversible termination sequencing processes.

(Left Panel) – For 454/Roche sequencing, the sample DNA is fragmented by nebulization and adaptors are linked to fragment ends. The DNA samples are then amplified by emulsion PCR. DNA-positive beads are then selected and placed in picolitre-sized wells and pyrosequenced. In the sequencing process nucleotide incorporation causes the release of inorganic pyrophosphate (PPi). The light emitted during PPi decomposition by the luciferase enzyme is then detected. The reaction is then repeated cyclically for each one of the four nucleotides (A,C,G,T). (Right Panel) – For Solexa/Illumina sequencing, the sample DNA is fragmented and adaptors attached to both ends of the sequences. The adaptor-DNA complex is then transferred to a solid flow cell channel where it is amplified by bridge amplification creating clusters of the same sequence with ~1000 copies. After amplification, the sample is denatured and the sequencing process starts. The adaptor primers, four fluorophore-labelled nucleotides and DNA polymerase are added in the first cycle. The nucleotides have their 3'OH group chemically blocked allowing only one nucleotide incorporation per sequence. The fluorophore is then excited by a laser and the incorporated base identified. After each cycle the blocked 3' terminus and the fluorophore are removed to allow a new detection cycle. The sequences length is dependent on the number of allowed cycles adapted from (Medini, Serruto et al. 2008) (Mardis 2008; Medini, Serruto et al. 2008; Pettersson, Lundeberg et al. 2009).

1.5 Sequence analysis tools

The work presented in this thesis was accomplished mostly through the use of custom computer code written in the PERL (Practical Extraction and Report Language) and R statistical computing languages. Besides these two languages, there are two other bioinformatics tools that were applied heavily in this thesis that are introduced here. These are a sequence similarity tool BLAT (Kent 2002) and the motif prediction tool Patser (Hertz and Stormo 1999).

1.5.1 Sequence similarity searches

Finding regions of similarity between pairs of sequences is a very common task in bioinformatics that is often a prerequisite to many other types of genome analysis. BLAT or the BLAST (Basic Local Alignment Search Tool) like alignment tool is an open source DNA and RNA alignment tool developed by Jim Kent (Kent 2002). As the name implies BLAT is similar to BLAST, in that both tools take short matches between two sequences and extend them in to High-Scoring Segment Pairs (HSPs) (Altschul, Gish et al. 1990; Kent 2002). BLAT was developed to be a faster version of BLAST that would be more suitable for analysis of large queries (Kent 2002). There are several differences between both of algorithms. For example, while BLAST creates an index of the query sequence, BLAT creates an index of the database (Altschul, Gish et al. 1990; Kent 2002). This speeds up the processing time when the database is of smaller size than the query sequences. BLAT starts with the indexing of every K-mer in the database and excludes the most common K-mer size words from the hash table. Second BLAST triggers an extension when one or two hits occur while BLAT can cause an extension on perfect or near perfect (one mismatch) matches of a defined number (default 11 bp for DNA sequences) (Altschul, Gish et al. 1990; Kent 2002). A sequence alignment can be as little as the size of the triggering K-mer that can vary between 8 and 12 bp. Thirdly, BLAST reports each area of homology independently while BLAT concatenates the homology blocks and reports them by target sequence. This is a particularly useful feature of BLAT if the analysis requires the sequential comparison of the same sequences to different databases, allowing comparison between the best hit in the first database with the best hit in the second database and identification of overlaps between them. The output from BLAT can be generated in 9 different easily parseable formats varying from the default psl format to BLAST-like (-m 8) tabular format.

1.5.2 Motif prediction

Many cellular processes are dependent on sequence-specific protein-DNA interactions. These interactions may be difficult to identify in primary sequences because of variability in the set of sequences bound by the protein. Because of this variability, when searching for instances of these interactions in sequences it is more suitable to use a pattern search tool than a sequence alignment tool. The Patser program is one such pattern search tool that allows for the identification of statistically significant matches between a model of protein-DNA specificity called a position weight matrix (PWM) and a set of target sequences (Hertz and Stormo 1999). A PWM is derived from a position frequency matrix (PFM) by calculating the natural logarithm of the frequency of each base scaled by the expected base composition at each position and base in the matrix. Patser iterates through each position in a target sequence and independently scores each position against the PWM. The P-value is then calculated as the probability of attaining an equal or higher score for each position (Staden 1989). The P-values and scores are therefore independent of the size of the sequence since each window is treated as independent observation and neighboring interactions are assumed not to occur. Each sequence will therefore have L-W+1 starting positions, with L representing the sequence length and W representing the alignment/matrix width (Hertz and Stormo 1999). The sequence length independency of calculated scores and P-values allows for the direct transference of significant P-values between two sets of sequences with different lengths (see chapter 3).

2 Testing the palindromic target site model for DNA transposon insertion using the *Drosophila melanogaster* P-element

2.1 Abstract

Understanding the molecular mechanisms that influence transposable element target site preferences is a fundamental challenge in functional and evolutionary genomics. Large-scale transposon insertion projects provide excellent material for studying target site preferences in the absence of confounding effects such as post-insertion evolutionary change. Growing evidence from a wide variety of prokaryotes and eukaryotes indicates that DNA transposons recognize staggered-cut palindromic TSMs. Here we use over 10,000 accurately mapped P-element insertions in the *Drosophila melanogaster* genome to test predictions of the palindromic target site model for DNA transposon insertion. We provide evidence that the P-element targets a 14 bp palindromic motif that can be identified at the primary sequence level, which predicts the local spacing, hotspots and strand orientation of P-element insertions. Intriguingly, we find that the although P-element destroys the complete 14 bp target site upon insertion, the terminal three nucleotides of the P-element inverted repeats complement and restore the original TSM, suggesting a mechanistic link between transposon target sites and their terminal inverted repeats. Finally, we discuss how the staggered-cut palindromic target site model can be used to assess the accuracy of genome mappings for annotated P-element insertions.

2.2 Introduction

Mobile DNA sequences known as TEs are naturally occurring mutagenic agents that have been harnessed as experimental tools for genetic analysis in a variety of model organisms (Craig 2002; Mates, Izsvak et al. 2007). Of the two major classes of transposable elements that exist – those that transpose directly via a DNA molecule (transposons), and those that transpose indirectly via a RNA intermediate (retrotransposons) (Craig 2002) – DNA-based transposons have been most widely developed as tools for gene disruption and gene transfer experiments, becoming essential parts of the genetic tool-kit in bacteria (Hutchison, Peterson et al. 1999), fungi (Ross-Macdonald, Coelho et al. 1999), plants (Kuromori, Hirayama et al. 2004) and animals (reviewed in Mates, Izsvak et al. 2007). One of the most advanced transposon systems for genetic analysis is the *Drosophila* P-element (Rubin, Kidwell et al. 1982), which has been engineered to facilitate a large number of genetic and genomic manipulations including gene disruption, reporter gene analysis, gene and enhancer trapping, misexpression of endogenous genes, and the generation of chromosomal aberrations (reviewed in Ryder and Russell 2003; see Chapter 1).

Because of the widespread utility of the P-element as a tool for *Drosophila* genetics and genomics, the mechanisms of P-element transposition have been studied intensively over the last 25 years (reviewed in Rio 2002; see Chapter 1). Like many DNA-based transposons, the P-element transposes through a "cut-and-paste" mechanism that can be divided into two events – excision from the donor site and insertion into a new location in the host genome. Transposition is initiated when the P-element encoded transposase protein forms a tetrameric complex that binds one of the P-element TIRs at the donor site (Beall and Rio 1998; Tang, Cecconi et al. 2007), followed by GTP-dependent synapsis with the other TIR and sequential cleavage of each TIR from the donor site (Tang, Cecconi et al. 2005; Tang, Cecconi et al. 2007). The P-element transposase complex then forms a staggered cut of 17-nucleotides at both TIRs (Beall and Rio 1997), exposing the reactive 3' single stranded extensions that mediate strand transfer and integration into a new target site (Beall and Rio 1998).

In contrast to donor excision and target site integration, the molecular mechanisms of target site selection for new P-element insertions remain poorly understood. Target site

selection at the genomic scale is generally thought to be non-random, with P-elements exhibiting a preference for insertion into euchromatic regions (Berg and Spradling 1991), a bias towards insertion into 5' end of genes (Kelley, Kidd et al. 1987), hotspots for insertion at both the gene (Green 1977) and nucleotide (O'Hare and Rubin 1983) levels, and local hopping in the vicinity of donor elements (Tower, Karpen et al. 1993). In addition to these factors that suggest the influence of chromatin structure, other studies have reported a role for local DNA sequence/structure in P-element target site selection.

Based on a limited sample of only 18 insertions, O'Hare and Rubin (O'Hare and Rubin 1983) first demonstrated that P-elements prefer to insert into an eight bp GC-rich consensus sequence (GGCCAGAC), which was later confirmed in a expanded sample (n=61 insertions) by Preston et al. (1996). Subsequently, many P-element insertion sites were shown to differ from this consensus sequence (Garrell and Modolell 1990), and other pre-genomic analyses of small samples led to different target motif (e.g. GXTCAGGC, Bellen, Kooyer et al. 1992), casting some doubt on the generality of the original target motif reported by O'Hare and Rubin (O'Hare and Rubin 1983). Liao et al. (2000) analyzed a much larger set of 1,469 P-element insertions sites mapped to partially assembled genome sequences and concluded that "although there are base preferences at each position, these are not strong enough to generate a clear consensus sequence." Instead these authors argued that the P-element recognizes a 14 bp palindromic structural motif based on a pattern of hydrogen-bonding at the target site. More recently, Julian (2003) analyzed a sample of 795 P-elements and reported a 14 bp non-palindromic consensus sequence (ANNGGCCAGACNNT) that extended the GC-rich motif of O'Hare and Rubin (1983). These conflicting results have lead us to clarify whether the P-element targets a specific motif and, if so, whether this motif is a palindrome in order to better understand the target site selection of the P-element and other DNA transposons.

The possibility that the P-element targets a palindromic motif is intriguing given the fact that many other DNA transposons in wide variety of organisms, including bacteria, plants, worms, insects and vertebrate, also appear to prefer palindromic target sequences (Table 2.1). A palindromic target site recognition model has potential relevance for understanding the mechanisms of transposon integration, since it is consistent with

transposase acting as multimeric complex with the target site DNA (Halling and Kleckner 1982; Davies and Hutchison 1995; Beall and Rio 1998; Hu and Derbyshire 1998; Haren, Ton-Hoang et al. 1999; Tang, Cecconi et al. 2007). Additionally, there may be functional connections between palindromic target sites and the TIRs that flank many transposons, which are themselves palindromic sequences. Finally, palindromic target sites are often observed for retroviruses (Wu, Li et al. 2005), which use integrase enzymes for integration that share catalytic activity with transposases (reviewed in Haren, Ton-Hoang et al. 1999). The palindromic nature of transposon target site recognition is not universally accepted, however, with both palindromic and non-palindromic TSMs often reported for the same transposon (cf. Korswagen, Durbin et al. 1996 and Preclin, Martin et al. 2003 for Tc1; see conflicting evidence for the P-element above). This uncertainty may have arisen because many pre-genomic analyses of transposon insertion site preferences were based on extremely small sample sizes of insertions, natural target sites that have undergone sequence evolution since transposon insertion, or insertions into small artificial target regions (e.g. plasmids) that only allowed a limited exploration of sequence space.

To understand transposon target site selection properly it is necessary to investigate a larger sample of target sites in their in vivo genomic context immediately following insertion. Large-scale transposon insertion projects, such as the P-element gene disruption projects in *D. melanogaster* (Spradling, Stern et al. 1995; Spradling, Stern et al. 1999; Hayashi, Ito et al. 2002; Bellen, Levis et al. 2004; Ryder, Blows et al. 2004; Thibault, Singer et al. 2004), provide excellent functional genomic data to study models of target site selection for DNA transposons. Here we analyze a sample of over 10,000 consistently mapped P-element insertions and provide evidence that the P-element prefers a staggered-cut palindromic target motif that can be identified at the primary sequence level. Moreover, we show that the local spacing, hotspots and strand orientation of P-element insertions across the genome support a palindromic insertion site model for transposon target site selection. These results have important implications for understanding the structure inverted repeat DNA transposons and their mechanisms of transposition, as well as for the analysis of artificial and natural transposon insertions in genome sequences.

Table 2.1 Palindromic transposon target site sequences are common across all major kingdoms of life.

Note that the length of the TSM is often longer than the target site duplication (TSD, indicated in bold). IUPAC ambiguity codes are as follows: N=A/C/G/T, W=A/T, Y=C/T, R=A/G, M=A/C, K=G/T.

Transposon	TSD (bp)	Target Site Motif (TSM)	Taxon	Reference
IS231A	11	**GGGNNNNNCCC**	Bacteria	(Hallet, Rezsohazy et al. 1994)
IS630	2	C**TA**G	Bacteria	(Tenzen and Ohtsubo 1991)
IS903	9	WTTYA**NNNNNNNNN**TRAAW	Bacteria	(Hu and Derbyshire 1998; Hu, Thompson et al. 2001)
Tn3/IS3000	5	TWN**TAWTA**NWA	Bacteria *	(Davies and Hutchison 1995; Kumar, Seringhaus et al. 2004; Seringhaus, Kumar et al. 2006)
Tn4652	5	G**TAWTA**C	Bacteria	(Kivistik, Kivisaar et al. 2007)
Tn5/IS50	9	A**GNTYWRANC**T	Bacteria	(Goryshin, Miller et al. 1998)
Tn10	9	GN**NGCTNAGCN**NC	Bacteria	(Halling and Kleckner 1982; Bender and Kleckner 1992)
Ac/Ds	8	**CTTATAAG**	Plant	(Kuromori, Hirayama et al. 2004; Ito, Motohashi et al. 2005)
Mu	9	CCT**NNNNNNNNN**AGG	Plant	(Dietrich, Cui et al. 2002; Fernandes, Dong et al. 2004)
Tc1	2	CAYA**TA**TRTG	Worm	(Korswagen, Durbin et al. 1996; Preclin, Martin et al. 2003)}
Tc3	2	AWA**TA**TWT	Worm	(Preclin, Martin et al. 2003)
Tc5	3	MY**TNA**RK	Worm	(Preclin, Martin et al. 2003)
Hermes	8	**GTGNNCAC**	Insect	(Guimond, Bideshi et al. 2003)
hobo	8	**GTTTAAAC**	Insect	(O'Brochta, Warren et al. 1994)
Minos	2	ATA**TA**TAT	Insect	(Metaxakis, Oehler et al. 2005)
Mos	2	AATA**TA**TATT	Insect **	(Granger, Martin et al. 2004)
P-element	8	ATR**GTCCGGAC**WAT	Insect	This study; (Liao, Rehm et al. 2000)
SB	2	RCAYA**TA**TRTGY	Vertebrate	(Vigdal, Kaufman et al. 2002; Carlson, Dupuy et al. 2003; Yant, Wu et al. 2005)

* Data are for a bacterial transposon mobilized in a fungal genomic background. ** Data are for an insect transposon mobilized in a worm genomic background.

2.3 Materials and Methods

P-element insertion sites were obtained from release 5.6 of the *D. melanogaster* genome annotation (Drysdale and Crosby 2005). The majority of these data are from large-scale transposon insertion projects (Spradling, Stern et al. 1995; Spradling, Stern et al. 1999; Hayashi, Ito et al. 2002; Bellen, Levis et al. 2004; Ryder, Blows et al. 2004; Thibault, Singer et al. 2004) with additional insertions curated from literature. Coordinates and strand information for the RS3 and RS5 P-element families were obtained from the DrosDel project (http://www.drosdel.org.uk/make_release5_GFF_inserts.php), since FlyBase release 5.6 does not contain information regarding the strand for the RS family of insertions. Data manipulation was conducted in custom PERL (version 5.8.6) programs using BioPERL (version 1.3) (Stajich, Block et al. 2002) modules. Data and statistical analysis was performed in the R programming language (version 2.6.2) (R 2008). In reality, P-element insertions occur between adjacent nucleotides in the genome and therefore should be annotated in genome sequences on inter-base coordinates. However, annotations in FlyBase are on base coordinates and therefore P-element insertion sites are represented differently on the positive and negative strands (i.e. at the base after the insertion site on the positive strand and at the base before the insertion site on the negative strand). To make coordinate systems comparable on the positive and negative strands for analysis of distances between P-element insertions, we added one base pair to the coordinates of insertions on the negative strand, but retain the annotated coordinate in Supplemental Files 2.1 and 2.2.

To determine if the P-element targets a specific motif at the primary sequence level, we generated sequence logos (Schneider and Stephens 1990) from sets of aligned P-element insertion sites. Insertions at the same coordinate on the same strand were collapsed to create sets of non-redundant insertion sites. To do this we extracted a 51 bp window centered around each insertion site (-25 and +25 from the insertion site) and used Weblogo (version 2.8.2) (Crooks, Hon et al. 2004) with the following options (c -k 1 -w 15 -h 5 -Y -B 0.5 -n -s -25 -T 0.1 -b). Logos were created for both positive and negative stand insertions for each "family" of P-elements generated from distinct insertion screens. Since sequence logos measure the information content and not the statistical significance of a motif, we tested each position in the motif for deviation from expected genome-wide base composition using a χ^2 test.

To measure the match of individual insertion sites to the putative P-element TSM, a PFM was generated from a non-redundant set of aligned P-element insertions sites. Insertions on the negative strand were reverse complemented before including into the initial PFM and scoring, thus all sites in our model are oriented relative to the positive strand. Since no significant differences were observed between nucleotide frequencies at complementary positions (e.g. positions one and 14; seven χ^2 tests, all $p>0.04$), we averaged frequencies of complementary nucleotides at corresponding positions around the plane of symmetry (e.g. positions one and 14) to construct our final PFM for scoring target sites. This palindromic PFM was used to score individual insertion sites using Patser (version 3b.5) (Hertz and Stormo 1999) with the following parameters: -A a:t 0.29 c:g 0.21 -d2 -R. For each insertion site, we evaluated the match to PFM by calculating a log-likelihood "motif score" for the distinct target sites that would give rise to that insertion site on the positive and negative strands. In addition, for each target site we calculated (i) a "half-site score" by assessing the match of the 5' and 3' half of the target site to the first seven columns of the 14 bp PFM, and (ii) a "palindrome score" that ranges from zero to seven, with a score of one given to each pair of corresponding positions in the palindrome that had complementary nucleotides and a score of zero given for non-complementary nucleotides.

2.4 Results

To ensure large enough sample sizes and reliable genome mappings for our analysis of P-element target site preferences, we restricted our analysis to four families of P-element (GT1, SUPor-P, EPgy2, and XP) from the *D. melanogaster* Release 5.6 genome annotation that were obtained from large-scale screens that were localized to precise sequence coordinates using inverse PCR after completion of the *D. melanogaster* genome sequence (Bellen, Levis et al. 2004; Thibault, Singer et al. 2004). These families of P-elements each had a large number of insertions (>500) with a high proportion of insertions mapped to a single base pair (>90%) and mapped to a specific strand (>90%). Preliminary analyses showed that inclusion of data from other P-element screens generated systematic biases in subsequent analyses because of conflicting genome mappings (see Discussion). Table 2.2 summarizes characteristics of genome mappings for 10,860 insertions from the four P-element families analyzed in this study.

Table 2.2 Summary of reliably mapped P-element insertions in release 5.6 of the Flybase genome annotation.

P-element family name	Total Number of Insertions	Number Mapped to 1 bp (% of Total)	Number Mapped to +/- Strand (% of Total)	Number on + Strand (% of Total)
GT1	556	531 (95.50%)	496 (93.4%)	260 (52.42%)
SUPor-P	2297	2288 (99.61%)	2134 (93.27%)	1065(49.91%)
EPgy2	3496	3473 (99.34%)	3258 (93.81%)	1630 (50.03%)
XP	5311	4974 (93.65%)	4972 (99.96%)	2479 (49.86%)
Total	11660	11266 (96.62%)	10860 (96.40%)	5434 (50.04%)

2.4.1 The P-element targets a 14- bp palindromic motif

We constructed separate sequence logos for insertions on the positive and negative strands for insertions that were mapped to a single base pair for each family. For all four families, we observed the same palindromic TSM for insertions mapped either to the positive or negative strands (Figure 2.1, Figure 2.2). The similarity in TSM for the different families suggests that the local target site preferences is intrinsic to the P-element and is not family or screen dependent. Therefore we pooled insertions for EPgy2, GT1, SUPor-P and XP into one sample for all subsequent the analyses of P-

element insertion preferences. These four families include a total of 10,860 insertions located in 10,221 non-redundant insertion sites in the *D. melanogaster* euchromatin.

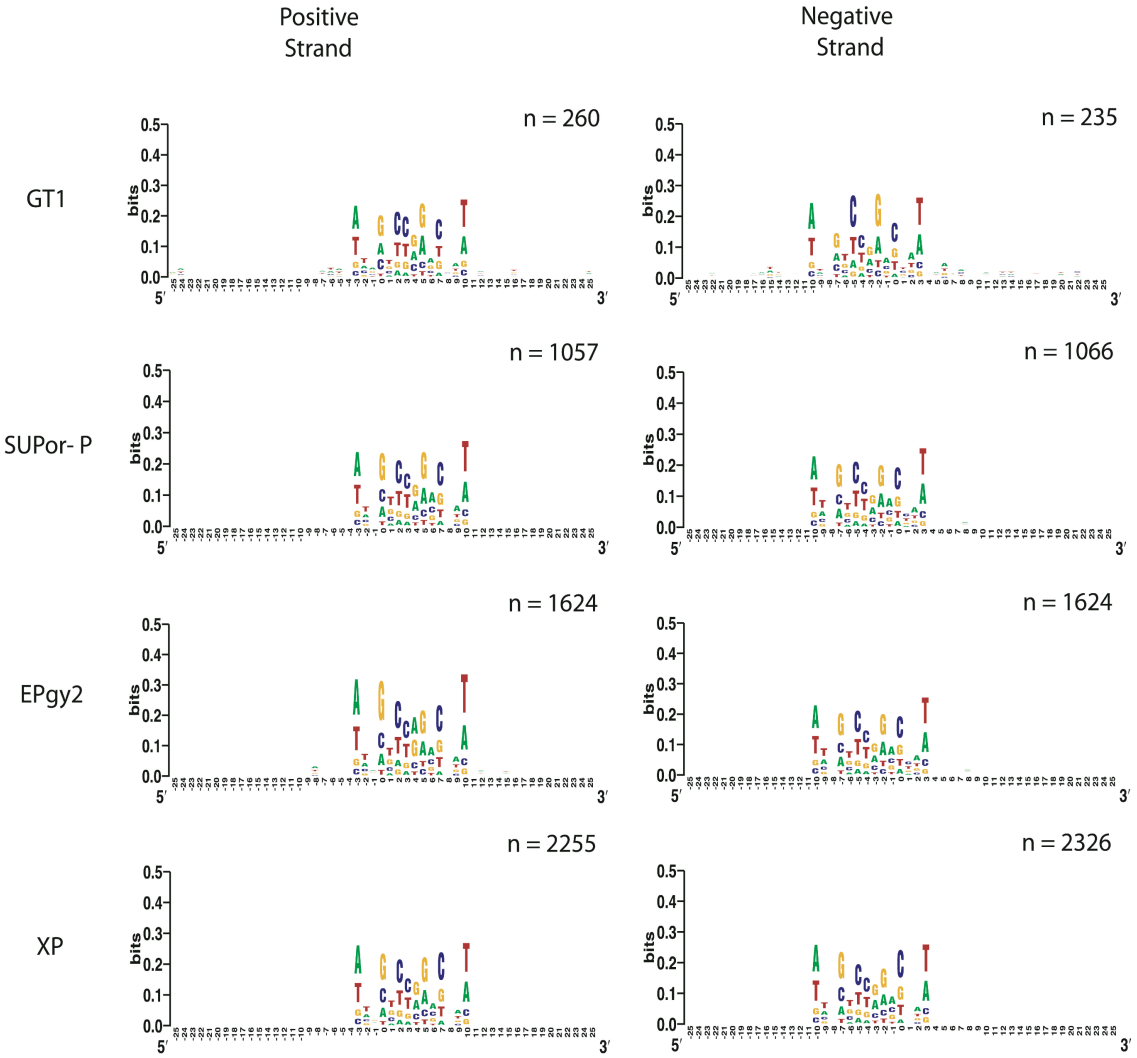
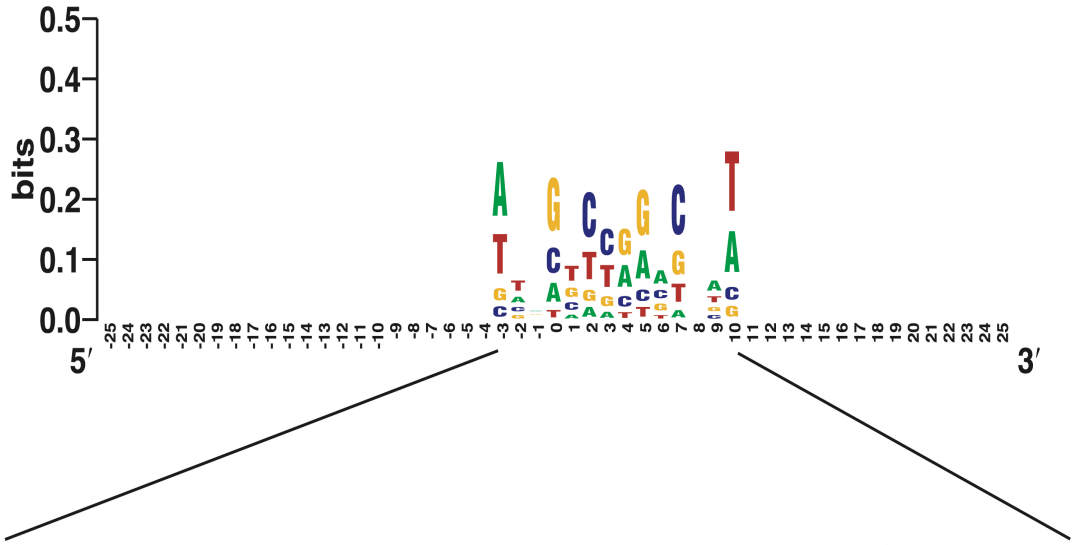


Figure 2.1 Sequence logos for the GT1, SUpor-P, EPgy2 and XP families.

Logos are arranged as in Table 2.2. Numbers reported are for the non-redundant set of insertions sites in the positive and negative strand used to construct the sequence logo. Note the numbers of non-redundant sites for each family do not sum to the total number of non-redundant insertion sites for the pooled dataset of all four families reported in the main text, since some insertion sites are found in multiple families and are made non-redundant in the pooled dataset.



	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10
A	**4618**	2489	**3004**	1933	1414	1221	1177	3389	3041	**3909**	918	2095	**3872**	3339
C	1015	2028	2242	2471	2220	**4654**	**3959**	1763	1327	2604	**4873**	**2868**	1809	1137
G	1166	1800	**2935**	**4934**	2615	1334	1680	**3947**	**4658**	2267	2474	2323	1952	983
T	3422	**3904**	2040	883	**3972**	3012	3405	1122	1195	1441	1956	**2935**	2588	**4762**
χ^2	2115	457	549	5347	1247	4173	2717	2726	4219	1174	5133	478	420	2323

Figure 2.2 The P-element targets a 14 bp palindromic TSM.

(A) Sequence logo depicting the relative base usage for a 51 bp window centered around 10,221 P-element insertion sites. The insertion site on the positive strand is just before position zero, and the insertion site on the negative strand is just after position seven. Insertions on the minus strand have been reverse complemented before being included in the alignment. The Y-axis is in bit (log base 2) units of the usage of bases in the motif relative to the random expectation of equal frequency. (B) Table of base usage in the 14 bp TSM and χ^2 statistics testing the null hypothesis that base usage at each position of the motif is random under the genome-wide background base composition in D. melanogaster. All positions deviate significantly from random base usage (3 degrees of freedom, $p < 2.2 \times 10^{-16}$ for all motif positions).

Alignment of these 10,221 high-quality P-element insertion sites in the *D. melanogaster* genome revealed an optimal 14 bp palindromic target motif with the consensus sequence ATRGTCCGGACWAT (Figure 2.2). This 14 bp palindromic TSM for the P-element is consistent with the 14 bp palindromic hydrogen bonding pattern reported for an independent set of insertions from the EP screen (Liao, Rehm et al. 2000), but differs from the originally reported 8 bp non-palindromic P-element TSM (GGCCAGAC, O'Hare and Rubin 1983). When oriented with respect to insertion sites on the positive strand, the center of the TSM is offset to the right of the insertion site (position 0), starting at position -3 and extending to position +10, since the P-element endonuclease makes a staggered cut with an 8 nucleotide 3' overhang upon integration. The central eight nucleotides of this motif represent the TSD generated by P-element upon

integration (O'Hare and Rubin 1983). The lowest information content positions in the motif directly flanking the core TSD base pairs where the P-element endonuclease cleaves DNA, and the highest information content site are at the termini of the motif (positions -3 and 10). In contrast to previous work (Liao, Rehm et al. 2000), we find strong statistical support for a clear consensus sequence: all columns in the 14 bp motif deviate significantly from the overall base composition of the *D. melanogaster* genome sequence (A=T=29%, G=C=21%; 14 χ^2 tests, 3 d.o.f., all $p<2.2x10^{-16}$) (Figure 2.2). We note that an important consequence of this staggered cut palindromic target site is that if P-elements are mapped to a single base pair consistently at either the beginning or end of the TSD, then each target site can lead to two distinct insertion sites, one each on the positive and negative strands.

2.4.2 The palindromic target site model predicts non-random local spacing of annotated P-element insertions

Under a model of a palindromic target site, we reasoned that if there are hotspot target sites in the genome into which multiple P-elements insert, they would integrate either in the same insertion site on the same strand, or into different insertion sites on opposite strands. Because P-element insertions are annotated to a single base pair, if such "opposite-strand" hotspot target sites exist in the genome, they are predicted to have a characteristic pattern of local spacing of eight bp distance between consecutive insertions, with one insertion on the positive strand followed by the next insertion on the negative strand. Figure 2.3 shows the distribution of distances between consecutive P-element insertions for all insertions, and for consecutive insertions on the same strand (+/+ and -/-) or opposite strand (+/- or -/+). The local spacing between P-element insertions shows a clear tendency for the P-element to insert with either a distance of zero or eight bp apart (Figure 2.3A). Consistent with the prediction of the palindromic target site model, the excess of zero bp distances are only found between consecutive insertions on the same strand (+/+ or -/-) (Figure 2.3B), while the excess of eight bp distances are found only between consecutive insertions on the +/- opposite strand configuration (Figure 2.3C) but not the -/+ opposite strand configuration (Figure 2.3D).

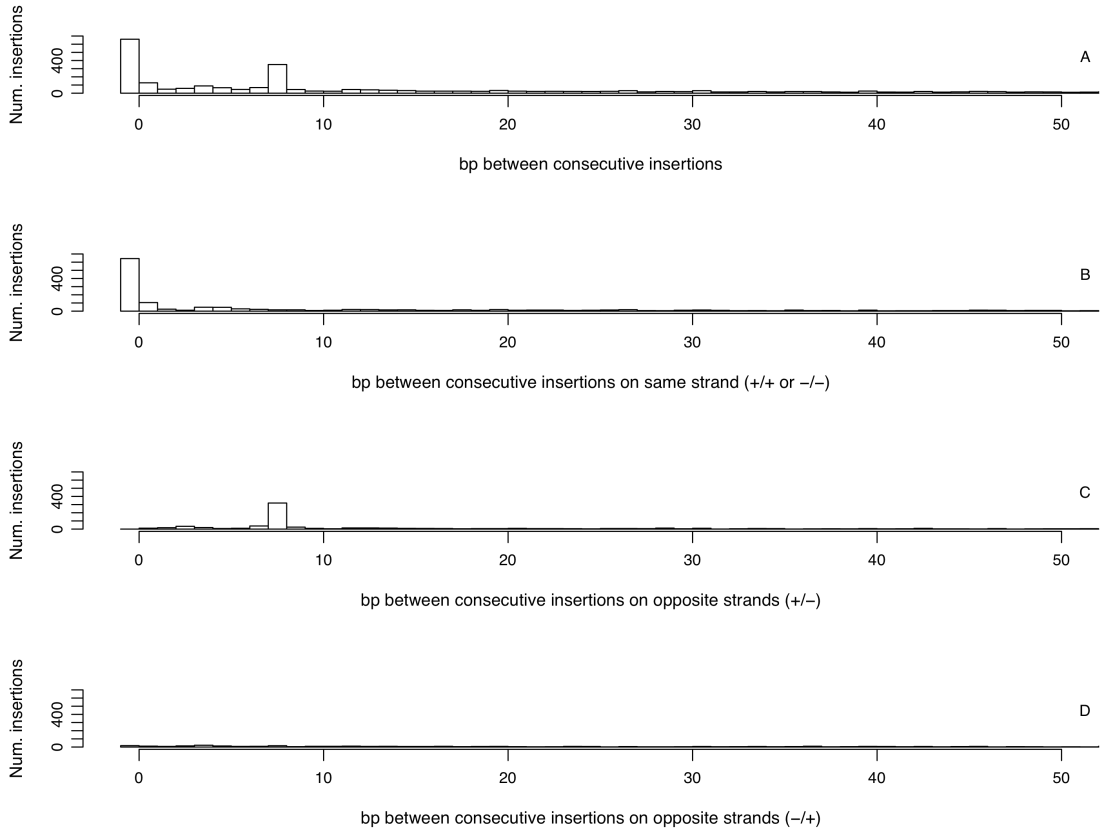


Figure 2.3 Non-random local spacing of P-element insertions mapped to a single nucleotide reveals two types of insertion hotspots.

(A) Distances, in base pairs (bp), between all consecutive P-element insertions in the genome. (B) Distances between consecutive P-element insertions on the same strand (+/+ or -/-), showing same-strand hotspots at a distance of zero bp. (C and D) Distances between consecutive P-element insertions on opposite strands (+/- or -/+), showing opposite-strand hotspots at a distance of eight bp. Note that the X-axis has been truncated at 50 bp in all three panels for clarity while the Y-axis goes up to 700 in all panels.

The excess of zero distances on the same strand is consistent with previous findings that the P-element often inserts into the exact same base pair in the genome (O'Hare and Rubin 1983; Roiha, Rubin et al. 1988; Bellen, Levis et al. 2004; Shilova, Garbuz et al. 2006). However, in contrast to previous reports that suggested insertion can occur in either strand at the same nucleotide (O'Hare and Rubin 1983; Roiha, Rubin et al. 1988), we find that the overwhelming majority (375/392, 95.6%) of insertion sites with more than one insertion at the same nucleotide occur on the same strand. Insertions into the same nucleotide on different strands in our data and previous reports most likely is due to the inconsistency in the placement of the P-element insertion at the 5' or 3' end of the TSD in different experiments. A tendency for P-element insertions annotated to a single

nucleotide to be spaced eight bp apart on opposite strands has not been reported previously, and is uniquely predicted under the 14 bp palindromic target site model for P-element integration, but not under a model of random integration.

These results also reveal that there are in fact two types of hotspot target sites when P-element insertions are mapped to a single nucleotide: (i) those that have multiple insertions into the same target site on the same strand characterized by a +/+ or -/- configuration a the same coordinate and (ii) those that have multiple insertions into the same target site on opposite strands characterized by a +/- configuration which is exactly eight bp apart. Moreover, the relative proportions of insertions into the two types of hotspot target sites (655 same-strand; 351 opposite-strand) are consistent with random strand integration, which are expected to occur in a 2:1 same-strand:opposite-strand ratio if the strand at a hotspot is chosen randomly (binomial test, $p=0.299$).

2.4.3 A palindromic target site model predicts hotspots for P-element insertion

If the palindromic motif in Figure 2.2 is a biologically meaningful representation of P-element target site preferences, we predict (i) that the observed P-element target sites should match the 14 bp motif better than background DNA sequences in the genome, and (ii) that hotspot target sites should match the 14 bp motif better than non-hotspot target sites. As found for the 14 bp hydrogen bonding pattern in (Liao, Rehm et al. 2000), P-element target sites have significantly higher scoring matches to the palindromic TSM relative to the distribution of scores for all possible target sites in the genome (Mann-Whitney U Test, $p<2.2x10^{-16}$) (Figure 2.4). We extend this finding to show that hotspot target sites for P-element insertion have better motif scores than non-hotspot target sites. This is true for all hotspot types: the 375 same-strand hotspot target sites, the 221 opposite-strand hotspots, and the 98 target sites that are hotspots by both criteria, all match the palindromic TSM better than the 9,208 target sites that are hit only once (Mann-Whitney U Tests, $p<2.2x10^{-16}$). In general, we observe that the rank order of median motif scores for the four classes of target sites is: non-hotspots < same-strand hotspots < opposite-strand hotspots < both-strand hotspots.

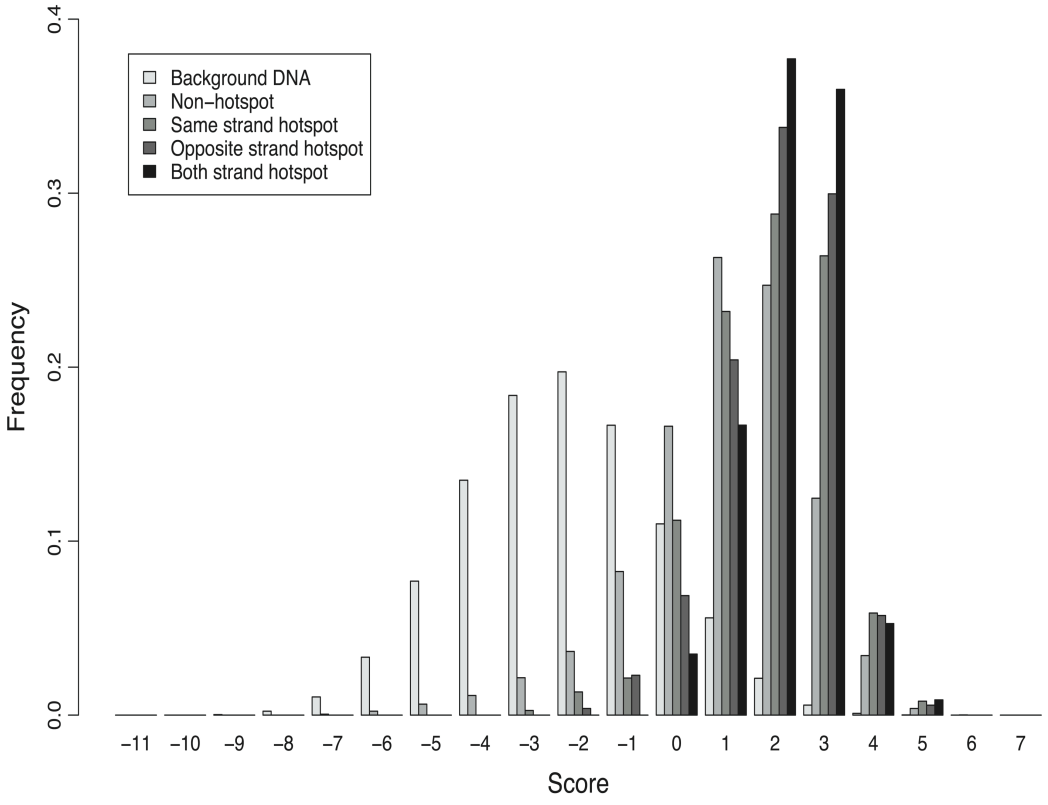


Figure 2.4 The 14 bp palindromic TSM discriminates P-element insertion sites, hotspots and background DNA.

Shown are the distributions of log likelihood scores of the 14 bp palindromic TSM relative to random background base composition.

The palindromic nature of the TSM raises the question of whether hotspots for P-element insertion might be influenced by whether a target site is simply a good palindrome or specifically a good match to the target site sequence. For example, if complementary substitutions occurred at corresponding positions (e.g. one and 14) of an optimal target site, the target site would remain a perfect palindrome but deviate by two substitutions from the optimal target motif. To evaluate whether hotspots are more influenced by match to the target sequence or "palindromicity", we tested for associations between the number of insertions per target site with motif score and/or palindrome score. In this analysis, we pooled all insertions from either same-strand and/or opposite-strand hotspots into the same target site giving a dataset of 9,902 non-redundant target sites. We found a highly significant positive correlation of number of hits per target site with motif score (Spearman's correlation, rho=0.154; $p<2.2x10^{-16}$) and weak positive correlation with palindrome score (Spearman's correlation, rho=0.029; p=0.003). We also evaluated the partial correlation of each score since motif

score and palindrome score are also positively correlated with each other (Spearman's correlation, rho=0.216; $p<2.2x10^{-16}$). This analysis revealed that the motif score given the palindrome score remains significantly associated with the number of hits per target site (Spearman's partial correlation, rho=0.151, $p<2.2x10^{-16}$), but not the converse (Spearman's partial correlation, rho=-0.0038; p=0.70). These results indicate that the match to the optimal target motif is more important in determining the frequency of P-element insertion than being a good palindrome.

2.4.4 No strand bias for P-element insertion

Because of the base pair complementarity of double-stranded DNA, matches to any palindromic motif should be distributed equally on both strands of the genome sequence, regardless of the motif sequence, genome-wide base composition or degree of mismatch allowed to the optimal motif. As expected under the palindromic insertion model, roughly equal proportions of P-elements insert into the positive and negative strands for all reliable mapped families of P-element (Table 2.2). Slight differences from the expected 50:50% ratio for a particular family are consistent with a small degree of experimental or computational error in strand mapping. Across all families, we find that 5,434 of the 10,860 (50.04%) P-element insertions that are mapped to a single base pair are found on the positive strand, which is not statistically different from the expected proportion of 50% (binomial test, p=0.9464). The lack of strand bias for the P-element is consistent with previous results showing that the distribution of insertion sites for the *C. elegans* Tc1 transposon is the same on the positive and negative strands (van Luenen, Colloms et al. 1994)

2.4.5 Evidence against sequential half-site recognition of palindromic target sites

As noted previously, matches to a palindromic motif score equally on both DNA strands, which raises the question: given a match to a full target site, how does the P-element determine which of the two possible strands to insert into if matches to the whole motif are equivalent on both strands? As has been suggested previously for other transposons (Halling and Kleckner 1982; Davies and Hutchison 1995; Hu and Derbyshire 1998; Rio 2002), the existence of a palindromic TSM for the P-element is consistent with the action of a multimeric transposase complex recognizing the target site. Biochemical evidence suggests that the P-element transposase acts in a tetrameric complex during donor excision (Beall and Rio 1998; Tang, Cecconi et al. 2007), and

therefore it is plausible that a multimeric complex is retained in the transposome during target integration. Under this model, we reasoned that the choice of strand might be mediated by sequential recognition of the half-sites by protomers of the multimeric transposome complex, which could lead to one strand of the DNA providing a better match to the seven bp half-site motif. For example, if the first protomer recognized the better half site, this could coordinate the transposome complex and lead to integration on the strand with the higher 5' half-site score. This scenario would allow for symmetry breaking of the full 14 bp palindromic target motif, and provide a mechanism for predictable strand selection. Since only 4^7 of all possible 4^{14} 14 bp sequences (0.006%, 1/16,384) are perfect palindromes, the vast majority of possible target sites breaks perfect palindrome symmetry and lead to a clear strand prediction.

To test if the half-site recognition model predicts the strand of P-element integration, we evaluated the difference in scores between the 5' and 3' half of each insertion site. We found no difference in the half-site scores that would support a model of half-site symmetry breaking through monomer recognition, with 49.9% (5,105/10,221) of insertion sites having a better 5' half-site score and 49.7% (5,090/10,221) having a better 3' half-site score, and only 0.2% (26/10,221) having equivalent 5' and 3' half-site scores. Thus, we conclude that the mechanism of P-element strand selection is inconsistent with a sequential half-site recognition model, but is consistent with simultaneous multimer recognition and random strand integration. The inability to find evidence for predictable strand integration supports the unbiased genome-wide strand mappings and the relative proportions of same-strand and opposite-strand hotspots reported above. Together these results are consistent with a model of random strand selection during target site integration, which parallels the random choice of which termini is chosen first in the P-element donor excision reaction (Tang, Cecconi et al. 2005).

2.5 Discussion

Understanding the mechanisms that control TE insertion and persistence in genomic DNA is a fundamental challenge in genome biology. Here we have used patterns of P-element insertion in the *D. melanogaster* genome to provide evidence for a staggered-cut palindromic target site recognition model for DNA transposon insertion, which has implications for both evolutionary and functional genomics. Consistent with another previous large-scale analyses (Liao, Rehm et al. 2000), we have found that the P-element targets a 14 bp palindromic TSM. We find evidence that the palindromic motif has a clear consensus sequence, whereas Liao *et al.* (Liao, Rehm et al. 2000) argued that it is a structural motif based on patterns of hydrogen bonds. As structure and sequence are intimately related at the DNA sequence level, we make no claim about which of these factors is causal. We have also shown that the local, non-random pattern of spacing between annotated P-elements is uniquely predicted by the palindromic TSM, and that matches to the TSM are a better predictor of P-element insertion frequency than palindromicity itself. We have further shown that there is no local or genome-wide strand bias for P-element insertion, consistent with a model of random strand integration. We conclude that the staggered-cut palindromic target site model is sufficient to explain the insertion preferences of the *D. melanogaster* P-element and, together with the widespread occurrence of staggered-cut palindromic target sites in disparate taxa, suggest that this model may apply generally to other cut-and-paste DNA transposons as well.

Our main findings are unlikely to be affected by systematic biases in our dataset since we have chosen to analyze large families of P-element insertions that have hallmarks of being accurately mapped to genome coordinates. However, some of our results, such as the relatively small difference in the score distribution of hotspot and non-hotspot target sites or the relatively low correlation of the motif score with the number of insertions per target site, can in part be explained by the fact that many P-element screens aimed to create non-redundant set of insertions for each gene in the genome (Spradling, Stern et al. 1995; Spradling, Stern et al. 1999; Bellen, Levis et al. 2004). Thus many additional target sites in our dataset are actually hotspots for P-element insertion but only have a single insertion. Despite this bias, the partition of all target sites into hotspot and non-hotspot sites is conservative with respect to the null hypothesis that there is no

difference between these categories in their similarity to the TSM. We also note that since the P-element TSM is constructed from a non-redundant set of insertions, an increase in the score of same-strand hotspots is not biased by multiple insertions at the same target site being represented multiply in the motif alignment. However, because opposite-strand hotspots were an unexpected result of our analysis, we did not consider these insertions as redundant in our original set of insertion sites. Thus insertions from opposite-strand hotspots are represented multiply in the motif alignment, and are expected to be biased towards higher match scores, as observed. Nevertheless, the same-strand hotspot results clearly demonstrate that the palindromic motif has explanatory power to discriminate P-element target sites from background and to discriminate hotspots from non-hotspot target sites.

2.5.1 Implications of the staggered-cut palindromic transposon target site model

Our analysis of the P-element target site preferences, along with a growing body of evidence from other DNA transposons (Table 2.1), suggests a general model for target site selection. The main feature of this model is that the optimal target site is a palindromic sequence/structural motif, which contains within it a staggered cut that is smaller than the length of the full target motif. This implies that sequences flanking the TSD are important for the target site selection (see also Bender and Kleckner 1992; Dietrich, Cui et al. 2002), that the target site is not the same as the TSD, and that each target site has two distinct insertion sites on the positive and negative strands when insertions are mapped to single nucleotide coordinates. Furthermore, since a palindromic motif is expected to be distributed equally on the positive and negative strand across the genome, DNA transposons are expected to insert with equal frequency on both strands. This last property of the palindromic target site model justifies the null hypothesis of studies that attempt to infer the post-insertion effects of natural selection from biases in transposon orientation in genome sequences (Smit 1999; Cutter, Good et al. 2005).

One key feature of the palindromic target site model we propose is that transposon integration destroys the original target site, leaving the TSD on both ends of the transposon, but only the 5' nucleotides flanking the TSM at the 5' end of the insertion, and *vice versa*. In the case of the P-element target site, the central eight bp are duplicated plus three bp of the target site nucleotides on either the 5' and 3' ends.

Intriguingly, we observe that the terminal three bp flanking the TSD at the 5' (ATR…) and 3' (…YAT) end of the TSM exhibit sequence complementary to the terminal three bp of the 3' (…ATG) and 5' (CAT…) ends, respectively, of the P-element TIRs (Figure 2.5A). Thus, although P-element integration destroys the complete 14 bp motif, the first three nucleotides of the TIR sequences inserted into the target site by the P-element complement the missing part of the target motif at both ends of the insertion. Because of the palindromic nature of the TIRs, complementation occurs regardless of whether the P-element is orientated in the 5' or 3' direction.

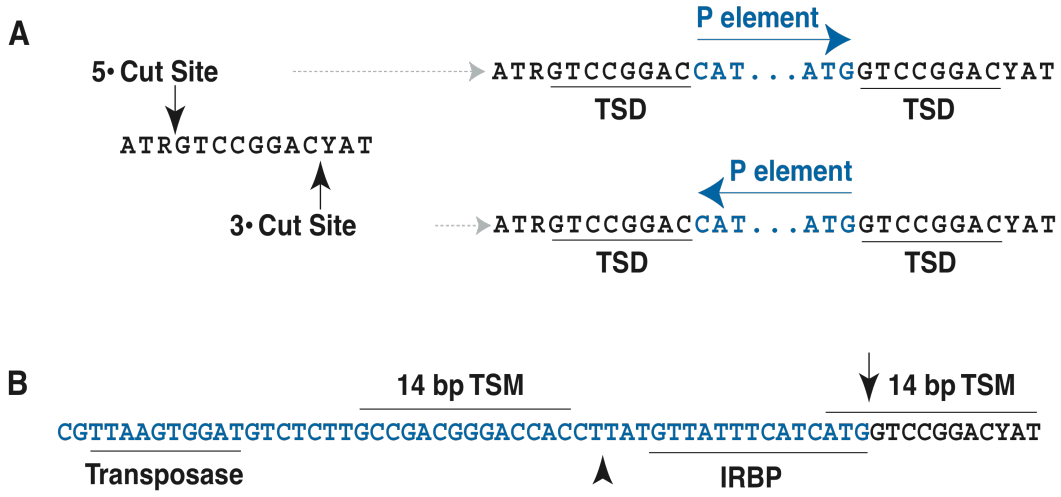


Figure 2.5 Model of P-element sequences in the context of the palindromic target site.

The P-element terminal inverted repeats complement the staggered-cut palindromic target site. Genomic sequences are shown in black, P-element sequences are shown in blue, and cut sites for transposase activity are shown as black arrowheads. (A) The terminal three nucleotides of the P-element inverted repeats restore and complement the optimal target sequences flanking the target site duplication (TSD). Note that this occurs on both ends of the P-element regardless of whether the 5' or 3' insertion site is used and the resulting orientation of the P-element insertion. (B) TSMs in the P-element terminal repeat and the target site flank the 17 bp staggered cut sites for donor excision. Shown also are the positions of binding sites for transpsosase and the inverted repeat binding protein (IR BP).

Complementation and restoration of the destroyed P-element target site suggests a mechanistic link between staggered-cut palindromic target sites and the structure of the TIR transposons, specifically involving the terminal nucleotides of the TIRs. In the case of the P-element, biochemical evidence shows a close association between the P-element transposase and the last two P-element nucleotides during donor excision (Beall and Rio 1998). Moreover, a special role for terminal nucleotides in the P-element TIRs may explain the strong conservation of only the first three nucleotides of the TIRs

among P-element family members in insects and vertebrates (Hammer, Strehl et al. 2005), and the widespread conservation of the first and last two nucleotides (5'-CA…TG-3') across diverse transposon families (Collins and Anderson 1994; Lee and Harshey 2003). The possibility that P-element sequences may complement and restore their target sites may also explain why P-elements continue to favor a target site even if when there is a pre-existing insertion (Tower, Karpen et al. 1993; Zhang and Spradling 1993; Timakov, Liu et al. 2002; Ryder, Blows et al. 2004), effectively allowing a non-lethal insertion to re-generate a "safe-haven" for other insertions. Moreover, since the P-element TIRs provide the optimal sequence for the restored TSM, P-element insertion is expected to improve the original TSM and make it more likely to be a hotspot. Finally, multiple insertions into the same target site are predicted to be in the inverted orientation and when insertions are mapped to base pair resolution separated by exactly eight bp, as has been demonstrated for the unstable singed^{weak} allele (Roiha, Rubin et al. 1988). The singed^{weak} allele is also hypermutable and undergoes reversion by precise excision of one P-element or the other at a high rate (Roiha, Rubin et al. 1988), and thus the recurrent targeting to safe-haven hotspots may ultimately increase subsequent rates of P-element remobilization.

The potential significance of target site complementation by the P-element termini is strengthened by the fact that high scoring sites for the TSM are found at positions 20-33 and 2875-2888 of the 5' and 3' P-element terminal inverted repeats, respectively, just internal to the inverted repeat binding protein site (Rio and Rubin 1988) (Figure 2.5B). These sites are in the upper 25th percentile of the distribution of target site scores, and are likely to be *bona fide* target sites since genetic evidence has revealed that a hotspot for P-element insertion exists at bp 19-26 of the P-element itself (Eggleston 1990). Since the P-element termini each carry one target site and the target site is duplicated and complemented at each end after insertion, four high scoring target sites (two at each end) are available for transposase activity at a donor site, which is fully consistent with the action of a tetrameric complex during donor excision (Beall and Rio 1998; Tang, Cecconi et al. 2007). Additionally, the two high scoring TSMs at each end of the integrated P-element closely flank the 17 bp staggered cut sites (Figure 2.5B), suggesting that the transposase complex may be coordinated to its cut sites during donor excision by the two TSMs.

More practically, the destruction of a full TSM on integration requires analysis of pre-integration (not post-integration) target sequences to determine true transposon target site preferences. Additionally, if terminal TIR sequences can partially complement post-integration target sites, it may be difficult to determine whether sequences at the termini of a single transposon insertion are part of the TIR or the target site. This issue was raised previously for the Tc3 transposon, and resolved by changing the sequences of target sites (van Luenen, Colloms et al. 1994). In fact, this ambiguity between TIRs and staggered-cut palindromic target sites may underscore the functional connections between the TIR structures of transposons and their target site sequences.

2.5.2 The palindromic target site model can be used to assess the quality of annotated transposon insertion sites

Although a palindromic target site model cannot predict the strand of a transposon insertion given a target site, it can be used to confirm the strand of an insertion site given its correctly annotated location in the genome. This is because under a staggered-cut palindromic model, transposons do not insert into the center of their target site. Therefore, an insertion annotated at a given nucleotide position in the genome should be generated by two different target sites on the positive and negative strand that can have different motif scores. To demonstrate the utility of this property of our model, we scored each P-element insertion site in our data set at the two potential target sites on either strand that would give rise to an insertion at this nucleotide to see if the higher scoring target site confirms the annotated strand in FlyBase. Remarkably, we found that the top-scoring strand under our palindromic motif model confirms the annotated strand for 90.4% (9,243/10,221) of P-element insertion sites in our dataset, confirming the high quality genome mappings of the four families analyzed here. The inability to perfectly confirm the annotated strand P-element insertion given its location is consistent with some residual error or inconsistency in the genome mappings in our dataset.

In contrast, we confirmed the strand for only 67.1% (3,823/5,694) of the remaining insertion sites mapped to a single base pair from other P-element screens omitted from our analysis (Supplemental File 2.2). We interpret this result to indicate that upwards of 20% of these P-elements from other families may have incorrect or inconsistent strand or coordinate mappings in *D. melanogaster* genome annotation, and this interpretation

is the primary reason these families were not analyzed here. These errors likely arise from multiple sources, as shown by differences in the sequence logos (Figure 2.6) and the rate of strand confirmation in the three most abundant P-element families not included in our study – EP, GawB, and LacW (Spradling, Stern et al. 1999; Hayashi, Ito et al. 2002; Bellen, Levis et al. 2004). For example, the GawB insertions show the correct sequence logo on the positive strand, but the logo appears to be shifted by one nucleotide on the negative strand, indicating a subtle difference in coordinate systems on the positive and negative strands. This is also reflected in the fact that we confirmed positive strand GawB insertions at the same rate (91%, 976/1,072) as accurately mapped families above, but we confirmed negative strand insertions at a much lower rate (66%, 681/1,031). In contrast, the EP family logos show much reduced information content, a shift in logos for both positive and negative strand insertions, and a lower rate of strand confirmation on the positive strand (53.6%, numbers) than on the negative strand (72.9%, 621/852). The lacW family appears to be the most poorly mapped set of insertions with nearly all insertions (86%, 543/1,012) mapped to the positive strand, logos that do not resemble any of the other families, and low rates of strand confirmation on both the positive and negative strands.

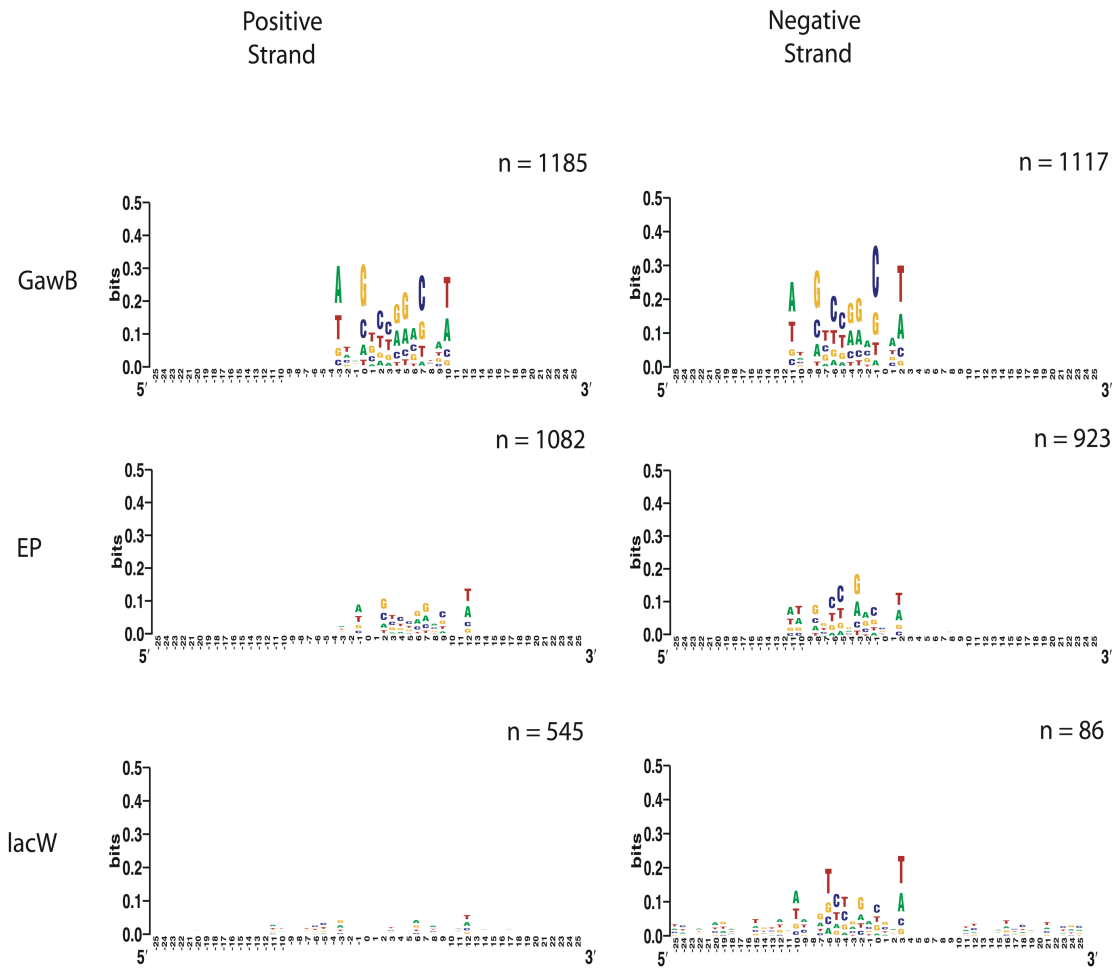


Figure 2.6 Sequence logos for the GawB, EP and lacW P-element families.

Numbers are for the non-redundant set of insertion sties on the positive and negative strand used to construct the sequence logo. We note that the total numbers of non-redundant insertion sites for each family include sites that have insertions for other P-element families and therefore are not equal to the number of family-specific insertions sites reported in the main text.

Finally, the DrosDel collection, with 3,194 insertions mapped to one nucleotide for the RS3 and RS5 types put together, also displayed unusual mappings. When we first analyzed these insertions independently from the remaining families we saw that although there was a strong logo (similar score to the one from the four families studied here) it was shifted to the left, starting at position -10 instead of position -3 (Figure 2.7A). In contrast with other families such as GawB and EP (Figure 2.6) the unusual logo for RS5/RS3 also did not seem to be strand biased (data not shown). Another factor that was peculiar about these insertions was the unusual low rate of correct strand prediction (783 of 2972, 26.35%) using the TSM derived shown in Figure 2.2. After correspondence with the DrosDel project team (Ed Ryder, pers. comm.) it became clear that these observations result from differences in the annotation process used by

DrosDel and the insertions in FlyBase that come from the GDP. The GDP consistently annotated P-element insertions to be at the genomic coordinate immediately 3' to the integrated P-element, which corresponds to the 5' side of the sequence that becomes the TSD after integration. In contrast, the DrosDel team used two different procedures during the annotation process. The first procedure annotated insertions at the 3' end of the element, which results in an identical annotation to the one used in most families in FlyBase (Figure 2.8A). The second procedure, which comprised the majority of the insertion sites in the DrosDel collection (78.57%), mapped the insertion to the 5' end of the element, which corresponds to the 3' side of the sequence in the genome that becomes the TSD (Figure 2.8B). To solve this inconsistency between GDP and DrosDel annotations, we shifted DrosDel insertion sites annotated at the 5' end 7 bp downstream on the negative strand and 7 bp upstream on the positive strand. After this process the logos (Figure 2.7B) and strand prediction rate (2693 of 2984, 90.25%) became similar to those in the four families used in this study, showing that differences in mapping procedures was the cause of the unusual sequence logos and low rate of strand prediction for the uncorrected DrosDel insertions.

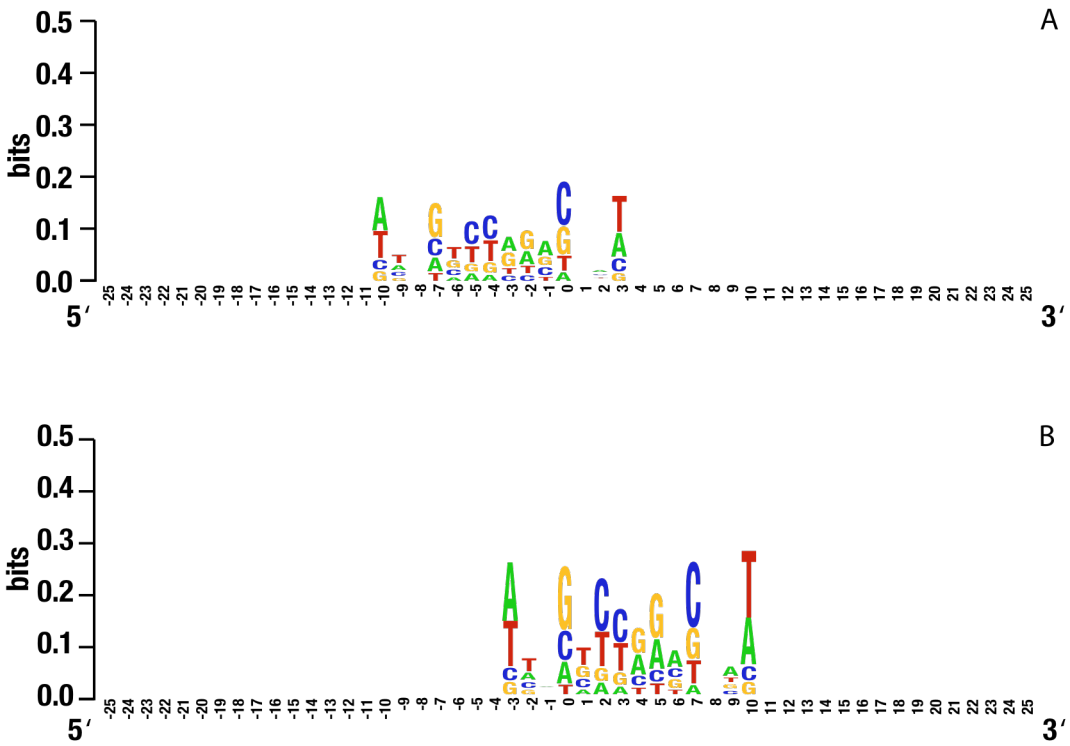


Figure 2.7 Target site motifs for the RS P-element family from the DrosDel project.

Logos constructed from DrosDel P-element insertions based on (A) 2,972 uncorrected insertion sites and (B) 2,984 corrected insertions sites. Before correction there were 3,229 insertions into 2,972 insertion sites with no strand ambiguity and 103 insertions with strand ambiguity totalling 3,332 insertions in 3,011 insertions sites. After correction there was 3,308 insertions into 2,984 non-ambiguous sites with 24 insertions into strand ambiguous sites totalling 3,332 insertions into 2992 insertion sites.

Together, these results show that our target site model can aid in the interpretation of the quality and consistency of annotated P-elements insertions. In some cases, such as GawB and the DrosDel insertions, we are able to suggest simple coordinate shifts that put the annotation of these families into register with the majority of P-element insertions in the *D. melanogaster* genome. For other families, however, the degree of these potential errors in coordinate or strand mapping are unknown but could have important consequences for use of these P-element collections by *Drosophila* researchers, including the misexpression of the incorrect neighboring locus for EP-elements mapped to the incorrect strand.

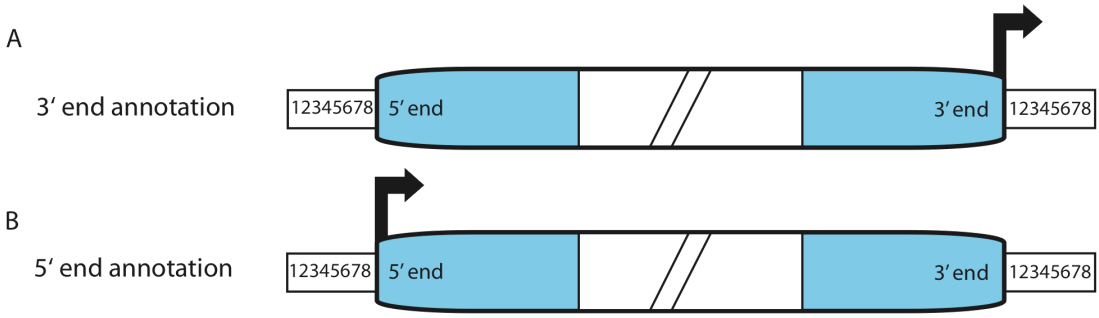


Figure 2.8 Different annotation procedures used the DrosDel RS elements.

Schematic of P-element insertion site mapping procedures at the 3' end (A) and 5' end (B). When an insertion is mapped to genome coordinates using the 3' end of the P-element, it will be mapped to the beginning of the TSD (12345678) (A). When an insertion is mapped using the 5' end of the P-element, it will be mapped to the end of the TSD. Differences in the end used for mapping procedures will therefore shift the TSM by the length of the TSD.

3 Promoter targeting preferences of the *D. melanogaster* P-element

3.1 Abstract

The *D. melanogaster* P-element is a DNA-based transposable element that is one of the most important tools in *Drosophila* genetics. One of the most striking features of the targeting preferences of the P-element is a strong tendency to insert into the proximal promoter regions of a subset of protein coding genes. Despite longstanding speculation about the role of open chromatin influencing the location of P-element insertion, the genomic determinants of P-element promoter targeting largely remain a mystery. By combining data from large-scale transposon insertion collections with sequence and chromatin properties of promoter regions, we attempt to decode the genomic factors associated with P-element promoter targeting in *D. melanogaster*. Our results indicate that the fine-scale distribution of P-element insertions in promoters is affected by avoidance of nucleosomes, but that nucleosome positioning is not the primary cause of promoter targeting. Investigation of other factors reveals that P-elements can insert into a wide variety of core promoter architectures, but prefer to insert in TATA-less promoters that have a paused RNA polymerase and are bound by specific members of the Trithorax and Polycomb groups of proteins. Our results provide the first genome-wide evidence that core promoter architecture and chromatin structure impact P-element target preferences, and shed light on the nuclear processes that influence the pattern of P-element insertion across the *D. melanogaster* genome.

3.2 Introduction

Transposable elements are a widespread and diverse set of mobile DNA sequences that influence many aspects of prokaryotic and eukaryotic genome biology. Based on the mechanisms by which they mobilize and insert genomic DNA, transposable elements can be broadly classified into DNA-based transposons and RNA-based retrotransposons (Craig 2002). Among the best studied of eukaryotic transposons is the *Drosophila melanogaster* P-element, which has had many aspects of its excision and integration processes characterized at the molecular level (reviewed in Rio 2002). This detailed understanding of the mechanisms of P-element transposition has allowed the development of an elaborate genetic toolkit in *D. melanogaster* using the P-element (Rio 2002). Nevertheless, many aspects of the P-element target preferences still remain unresolved, and have hindered the use of this transposon in generating insertion mutants in all *D. melanogaster* genes (Spradling, Stern et al. 1999; Bellen, Levis et al. 2004; Thibault, Singer et al. 2004).

The most unique feature of the P-element is its preference to insert into the proximal promoter regions and 5' untranslated regions (UTRs) of protein coding genes (Tsubota, Ashburner et al. 1985; Searles, Greenleaf et al. 1986; Kelley, Kidd et al. 1987; Spradling, Stern et al. 1995; Bellen, Levis et al. 2004). This strong non-random promoter targeting is thought to arise from the increased accessibility of the P-element transposase to genomic DNA resulting from open chromatin during transcription (Kelley, Kidd et al. 1987). While little direct evidence has been provided in support of this hypothesis (but see Voelker, Graves et al. 1990), many P-elements clearly do insert into regions of active chromatin, fortuitously allowing for the analysis of endogenous gene expression patterns through enhancer trapping (Bellen 1999). A second well-known feature of the P-element is that it only targets a subset of ~40% of *Drosophila* genes, even when insertion mutants are generated in very large numbers (Bellen, Levis et al. 2004). A corollary of this property is that some target genes are known to be hotspots for P-element insertion, such as the singed gene (Roiha, Rubin et al. 1988). Previous reports have suggested that these target gene preferences arise from the P-element inserting into genomic regions that are active during male germline development (Bownes 1990; Timakov, Liu et al. 2002) or genes with "poised" promoters that have an open chromatin structure and paused RNA polymerase, such as

the *Heat shock protein* (*Hsp*) genes (Lerman, Michalak et al. 2003; Walser, Chen et al. 2006). However, the biological basis for why the P-element prefers to insert into promoter regions and only into a subset of host genes largely remains a mystery.

One potential explanation for both the promoter targeting and non-random target gene preferences is that the P-element recognizes some aspect of core promoter architecture present in a subset of *D. melanogaster* genes. Core promoter architecture is known to vary across the *D. melanogaster* genome at the sequence level, with many known and unknown promoter motifs generating different subclasses of promoter (Ohler, Liao et al. 2002; FitzGerald, Sturgill et al. 2006). The best-studied promoter motifs in *Drosophila* are the TATA box motif, the initiator (Inr) motif and the downstream promoter element (DPE). The TATA box can recruit general transcription factors (GTFs) by itself (Juven-Gershon, Hsu et al. 2006) and is often found in an AT-rich region from -30 to -20 nucleotides upstream from the transcription start site (TSS) (Kutach and Kadonaga 2000). In *D. melanogaster*, the DPE sequence appears to be as common as the TATA box promoter (Kutach and Kadonaga 2000), with a core 6 bp sequence often found between +28 and +33 nucleotide downstream of the TSS (Kutach and Kadonaga 2000; Lim, Santoso et al. 2004). The DPE is unable to recruit GTFs by itself, and instead works in synergy with the Inr with a characteristic spacing between these motifs (Kutach and Kadonaga 2000). Core promoter architecture has been shown to vary according to gene structure (Zhu and Halfon 2009) and gene expression (FitzGerald, Sturgill et al. 2006). Likewise, specific core promoter architectures are associated with alternative mechanisms of polymerase recruitment (Isogai, Keles et al. 2007) and elongation (Hendrix, Hong et al. 2008). Thus, there are many plausible developmental and genomic mechanisms as to how variation in core promoter architecture may impact the accessibility of transposon insertions in promoter regions.

Here we aim to test whether core promoter architecture impacts the target preferences of the *D. melanogaster* P-element. To address this question, we studied the sequence and chromatin factors of promoters associated with induced P-element insertions in *D. melanogaster*. Specifically, we investigate how the presence or absence of P-elements in a promoter region is affected by nucleosome occupancy, RNA polymerase II pausing, core promoter motif composition, basal transcription factors and chromatin remodeling factors. We find that P-elements can insert into a wide variety of promoter architectures,

but prefer to insert in TATA-less, paused promoters bound by the TBP-related factor 2 (TRF2) that contain DNA replication element (DRE) motifs and H3K4me3 modified nucleosomes. We also link the fine-scale spatial distribution of P-element insertion in promoter regions to nucleosome positioning. Together our results provide the first genome-wide evidence that core promoter architecture and chromatin structure directly influence P-element target preferences, and shed light on the nuclear processes that constrain the pattern of P-element insertion across the *D. melanogaster* genome.

3.3 Materials and Methods

P-element insertion sites and mRNA annotations were obtained from release 5.14 of the *D. melanogaster* genome from FlyBase. Based on previous results (Linheiro and Bergman 2008), we only analyzed P-element insertion site data from 5 reliably mapped families: GT1, SuPor-P, EPgy2, XP and GawB. GawB insertions on the negative strand were shifted by +1 bp to account for systematic differences in the mapping of this family. We excluded all P-element insertions from these families that were not mapped to a single base pair or did not have an annotated strand. In total, we analyzed 13,346 P-element insertions in 12,267 non-redundant insertion sites.

A dataset of 14,229 non-redundant TSSs was extracted from the release 5.14 annotation by parsing all non-redundant start coordinates of mRNAs from protein coding genes that use RNA polymerase II promoters. If a gene did not have an mRNA transcript (such as for tRNA, 5SrRNA, snoRNA and snmRNA genes), if an mRNA transcript did not have a defined strand, or if an mRNA transcript had no UTR (e.g. the mRNA start site is the same as its CDS start site) it was discarded. Genes on the negative strand were reverse complemented so all TSSs were processed in the same orientation.

To find predicted promoter motifs, sequences flanking TSS (-250, +250) were scanned with Patser (version 3b.5) using position weight matrices from Ohler et al. (2002), JASPAR (Sandelin, Alkema et al. 2004) and Gershenzon et al. (2006). Motif score cut-offs were determined by evaluating motif predictions against the curated data set of TATA and DPE motifs from Kutach and Kadonaga (2000). To do this we extracted promoter sequences from the *Drosophila* core promoter database (DCPD), mapped them to the *D. melanogaster* genome sequence using BLAT (version 34) and migrated the local coordinates of annotated promoter motifs in DCPD to Release 5 genome coordinates (Supplemental File 3.1). Final parameters for motif annotation were chosen based on receiver operator characteristic (ROC) analysis between the validated promoter motifs from DCPD and our promoter motif predictions (Supplemental File 3.2). To assess how robust our results are to independent promoter motif annotation methods, we also repeated our analysis using promoter motif annotations from Zhu and Halfon (2009) for 12,588 of their TSSs that mapped exactly to our TSSs (7,917 TSS greater than 2 Kb apart).

Chromatin immunoprecipitation (ChIP) data for the status of RNA polymerase II in TSSs was obtained from two sources (Muse, Gilchrist et al. 2007; Zeitlinger, Stark et al. 2007) and linked to our data by matching their mRNA CG ids to the CG ids for the mRNAs in our dataset of TSSs. In total, we were able to match 9,310 CG ids from (Muse, Gilchrist et al. 2007) and 9,077 CG ids from (Zeitlinger, Stark et al. 2007). Chromatin IP data from GTFs (Isogai, Keles et al. 2007), nucleosome occupancy (Mavrich, Jiang et al. 2008), and *Polycomb* and *trithorax* protein groups (Schuettengruber, Ganapathi et al. 2009) were cross-referenced to our TSSs using overlapSelect (version 211) from University of California Santa Cruz (UCSC) genome browser tools with a window of ±1Kb from the TSS (see below).

We performed preliminary investigations of the univariate effects of each genomic factor (e.g. RNA polymerase) on P-element occupancy by calculating an enrichment factor, and assessing statistical association using χ^2 tests of independence. Where multiple datasets for a given factor were available (e.g. RNA polymerase, motif predictions), we assessed the dependency of the dataset used on the association between a genomic factor and P-element insertion. We then selected the most relevant dataset to include in a multivariate generalized linear model (GLM) with binomial errors to assess interactions among factors. We coded the lack of a genomic feature as the default level in the GLM. We then performed model reduction to eliminate unnecessary factors and interactions.

All data manipulation was conducted with custom PERL (version 5.8.6) programs using BioPERL (version 1.3) modules (Stajich, Block et al. 2002). Graphical and statistical analysis was performed in the R programming language (version 2.9.1) (R 2009).

3.4 Results

3.4.1 Evaluation of Patser-based promoter motif predictions

Since there is not an official annotation of promoter motifs in *Drosophila*, we developed a motif prediction approach using Patser to annotate each TSS for core promoter motifs and evaluated this approach using curated data from the DCPD. The DCPD annotates the TATA and DPE motif for 205 *D. melanogaster* promoter sequences. The INR is assumed to be present in all DCPD promoters at the TSS, and therefore we cannot train INR prediction methods on these data. We mapped 191 of these promoters to a unique position in the *D. melanogaster* genome sequence (14 mapped to multiple locations), and retained 159 promoters that overlapped a FlyBase TSS within ±250 bp (32 did not overlap a FlyBase TSS), since these are most likely to represent *bona fide* TSSs. In these 159 promoters, 67 contained an annotated TATA motif and 64 contained an annotated DPE motif (Supplemental File 3.2) that we could use to train our motif prediction methods.

We first tested how well our promoter motif annotation strategy identified known TATA and DPE motifs in the DCPD annotation using PWMs from three different sources (Ohler, Liao et al. 2002; Sandelin, Alkema et al. 2004; Gershenzon, Trifonov et al. 2006) with search windows specified by Ohler et al. (2002) (TATA: -60 to -15 and DPE +10 to +27). We identified PWM score cut-offs that showed the best ROC graph performance. Using these windows, the TATA motif could be predicted accurately for all the 3 TATA PWMs, with an area under the ROC graph curve above 94% (Figure 3.1). The results for the DPE motif were less encouraging with areas under the ROC curve ranging from 60% to 65%. Therefore, we tested several other windows and score cutoffs for the DPE motif starting with the predictions from (Ohler, Liao et al. 2002) until we could not improve the ROC performance score anymore. We arrived at an improved ROC performance through more restrictive windows (TATA: -50 to -15 and DPE at positions +17, +25 and +27) (Figure 3.1). For the DPE motif the JASPAR motif presented the best results, with an area under ROC curve ~96%, in comparison with the other motifs (Ohler: 82% and Gershenzon: 88%). We note that the higher performance of the JASPAR PWMs is expected since DCPD sequences were used to generate the JASPAR PWMs.

Since the optimal window we found for the DPE motif prediction is only one bp long, it is essential to the get the location for the DPE motif in our methods correct, so that we can confidently transfer results from our training set to the entire *D. melanogaster* genome. Therefore, we tested whether the optimal location of the DPE motif in the DCPD training set is the same as our genome-wide dataset of TSS. To do this, we plotted DPE motif matches in the -60 to +60 window for all PWM matches that were below the P-value of the predicted best score for the genome-wide set of TSSs and for the DCPD TSSs. As shown in Figure 3.2, the DCPD and genome-wide set of TSSs show peaks in the DPE motifs at the same locations, suggesting our training set is representative of the entire genome. Thus, the final motif prediction parameters used here was based on the adjusted windows and cutoffs from the best results from the ROC analysis for both the DPE and TATA.

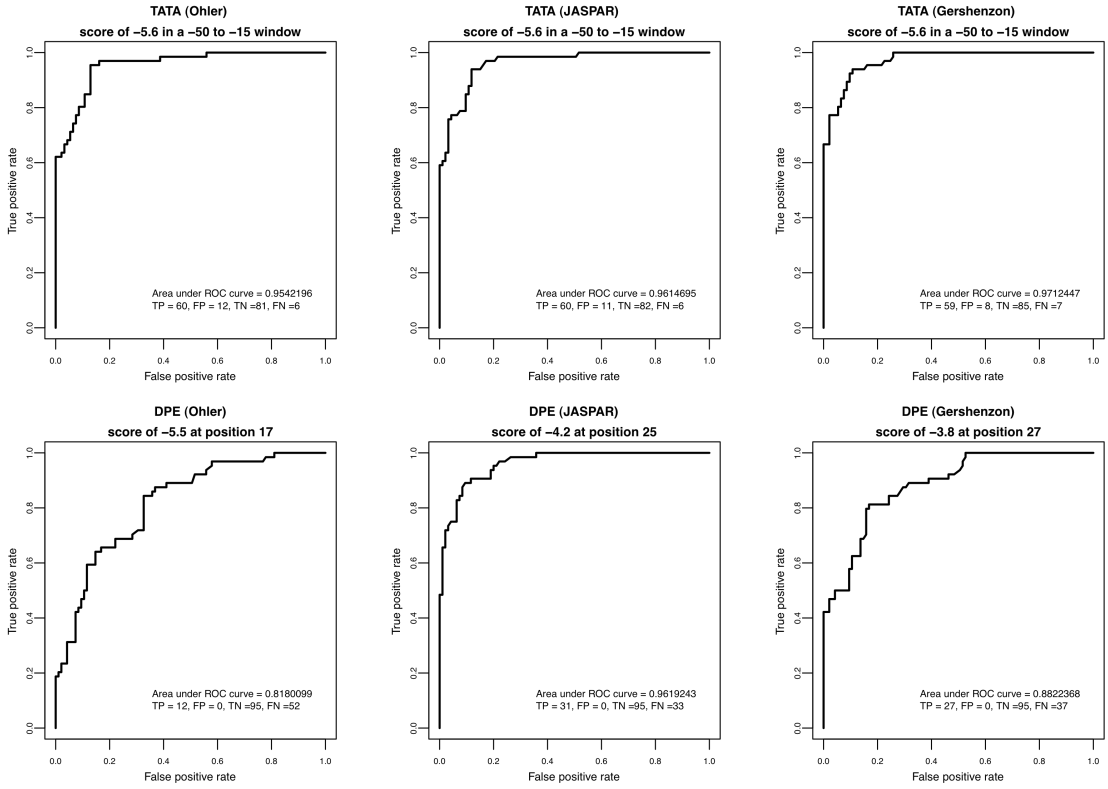


Figure 3.1 Performance of Patser-based promoter motif prediction with different PWMs

ROC performance graphs for the comparison of our promoter motif predictions with the annotated promoters from DCPD based on the ranking of the score/natural logarithm of the P-value for the matches. On top of each graph, the source of the PWM, the score cut-off and the window used are indicated. Inside the graphs the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) are indicated alongside with the area under the ROC curve. The top 3 graphs are for the TATA motif predictions that have the same windows (-50 to -15 bp from the TSS) and score cut-offs (below -5.6). The bottom three graphs for the DPE motifs showed more variation between PWMs, with optimal cutoff scores ranging from -3.8 for the PWM from Gershenzon et al. (2006) to -5.5 for the PWM from Ohler et al. (2002) and with the DPE location also varying from 17 for the Ohler et al. (2002) PWM to 27 for the Gershenzon et al. (2006) PWM. Recent results have shown that the DPE PWM from Ohler et al. (2002) was mixed with the motif ten element (MTE) (Rach, Yuan et al. 2009), explaining the bigger shift in the optimal location for DPE in relation to the other 2 matrixes.

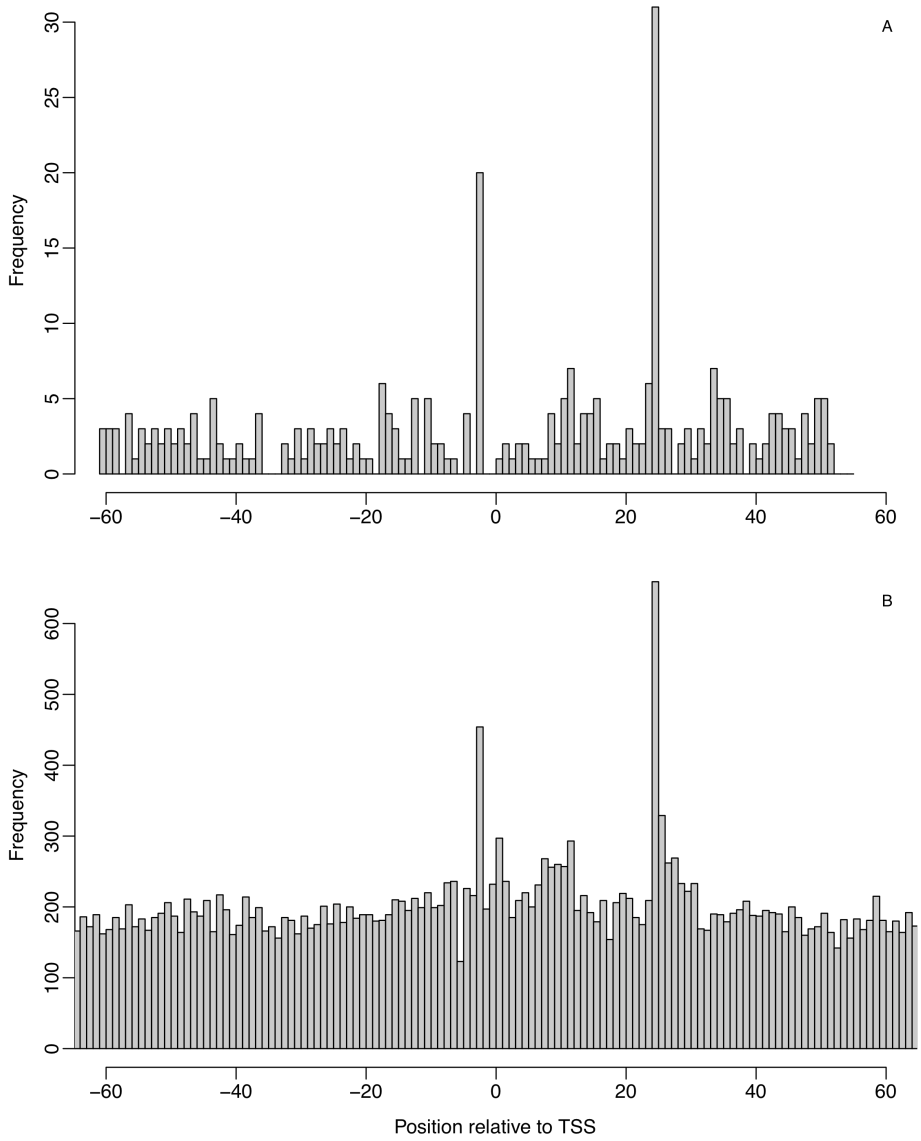


Figure 3.2 Location of predicted DPE motifs in TSSs from the DCPD and the entire *D. melanogaster* genome.

Our TSS annotation procedure considers the TSS as bp 0 while the annotation in DCPD considers the TSS at bp +1. After compensating for this discrepancy, we performed motif predictions and plotted the frequencies for each location bellow the predicted score to see if the DPE motifs in the DCPD (A) and genome-wide dataset of TSSs (B) were equivalent. For both datasets, windows with scores below -4.2 for matches to the DPE matrix from the JASPAR database are plotted. Two distinct peaks are visible at positions -2 corresponding to the Inr motif and another at position +25 corresponding to the DPE motif. The peak at the Inr position is due to the similarity between the DPE matrix and the Inr motif and the co-occurrence of both motifs.

3.4.2 Promoter targeting of P-element extends ±1000 bp from the TSS

Although it is well established that P-elements prefer to insert near the 5' end of protein coding genes (Tsubota, Ashburner et al. 1985; Searles, Greenleaf et al. 1986; Kelley, Kidd et al. 1987; Spradling, Stern et al. 1995), the scale of this promoter targeting preference has only been investigated in limited detail. Bellen et al. (2004) reported that "a large fraction of all P-element insertions associated with genes fall within 500 bp of the transcript start site." To quantify the scale of the P-element promoter targeting in better detail, and to establish criteria for classifying whether a promoter region was associated with a P-element or not, we plotted the number of P-element insertions and non-redundant P-element insertion sites in 100 bp windows from ±100 to ±5000 bp around the TSS (Figure 3.3A). For window sizes smaller than ±1000 bp from the TSS, there is a non-linear, increasing relationship between distance to the TSS and both the number of P-element insertions and the number of P-element insertion sites. Beyond ±1000 bp from the TSS, the presence of a P-element increases linearly with distance from the TSS, suggesting random probability of insertion with increasing target size at this scale. These results indicate that the promoter targeting preferences of the P-element spans a scale of ~1000 bp on either side of the TSSs.

One potentially confounding effect of using this relatively large window size to classify promoters according to their P-element occupancy is that neighboring TSSs from the same or different gene can fall within the ±1000 bp range. Indeed the majority of TSSs in the *D. melanogaster* genome are less than 1000 bp from their nearest neighbor (Figure 3.3B). Thus, to avoid the effects of redundancy of counting P-element insertions that are contained in more than one promoter region as defined by our ±1000 bp window around the TSS, we also analyzed a subset of TSSs in the genome that are greater than 2000 bp from their nearest neighbor. The same spatial scale of P-element promoter targeting (±1000 bp) is observed in this non-redundant dataset (Figure 3.3A; Figure 3.4A). Based on these results, we used a window of ±1000 bp around the TSS to classify promoters with respect to the presence or absence of P-element insertions in the following analyses. Using these criteria, we classify 6,005 TSS with P-elements and 8,224 TSS without P-elements across the entire genome. Likewise, we classify 3,243 TSS with P-elements and 5,481 TSS without P-elements for TSS that are greater than 2 Kb away from any other TSS in the genome. Genome-wide we find 9,324 P-element

insertions in 8,446 non-redundant insertion sites in all promoter regions, and we find 8,148 P-element insertions in 7,343 non-redundant insertion sites in promoters with TSS greater than 2 Kb from their nearest neighbor.

3.4.3 P-elements orient randomly with respect to the direction of transcription.

We have previously shown that the P-element targets a palindromic insertion site motif that leads to random integration in to the positive and negative strand, both locally within individual targets sites and globally across the genome (Linheiro and Bergman 2008) (Chapter 2). It is possible, however that the orientation of P-element insertions in promoter regions may be influenced the orientation of the transcription unit of the target gene that we would not have detected in our previous analysis. Consistent with random choice of strand within individual targets sites or across the entire genome, we find that the P-element inserts randomly in promoters with respect to the direction of transcription of the gene into which it inserts (Table 3.1). This is true for the number of insertions or insertions sites, upstream or downstream of the TSS, for all TSSs and those greater than 2 Kb from each other (Binomial tests, all $P>0.03$). Thus, we pooled counts for all P-elements insertions on both strands within a promoter region in all further analyses.

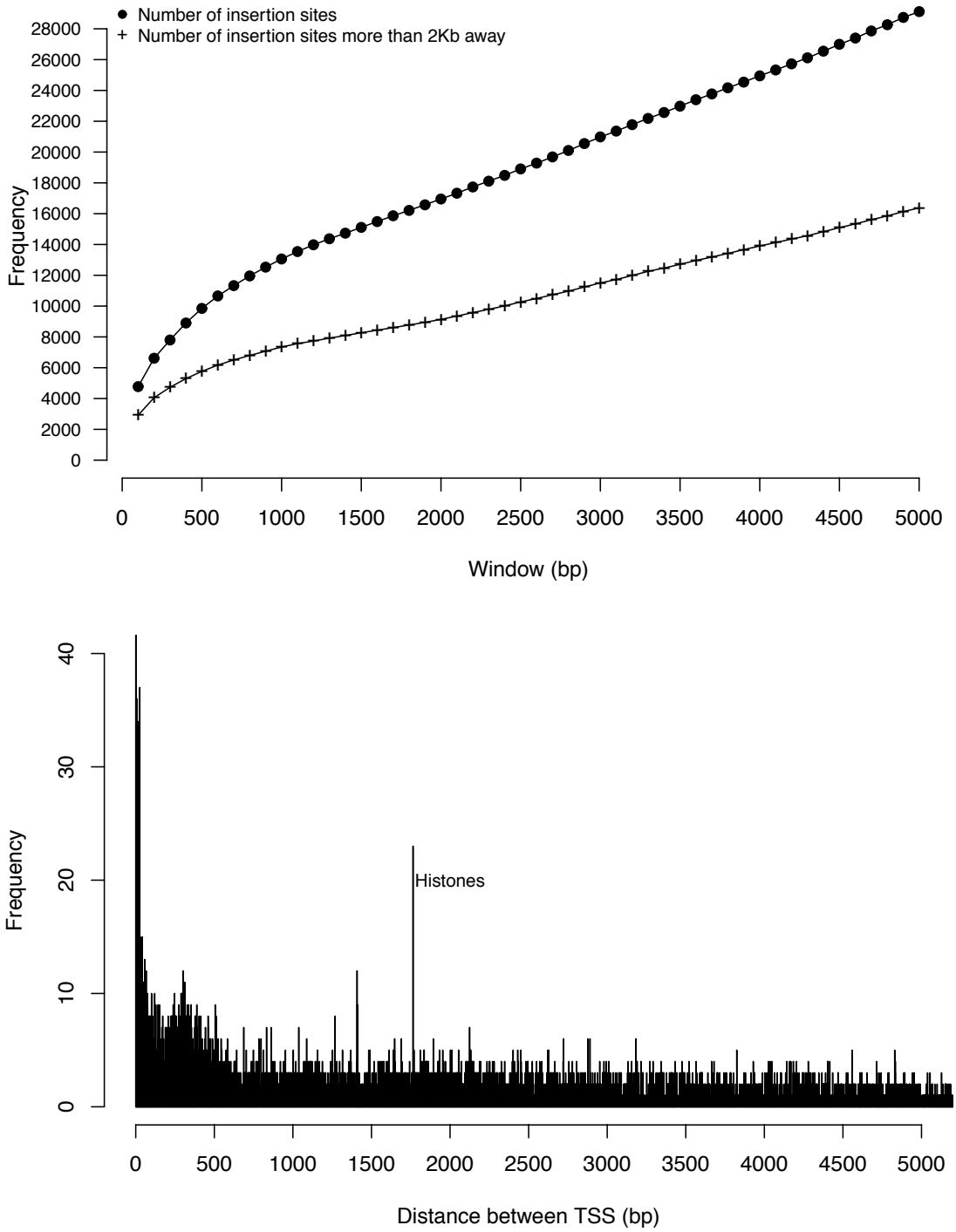


Figure 3.3 Distance between P-elements and TSSs and between adjacent TSSs

(A) Cumulative frequency of non-redundant insertion sites around the TSS in windows increasing by 100 bp from the TSS on both 5' and 3' sides, ranging from ±100 to ±5000 bp. The two lines show that TSSs more then 2 Kb apart from each other (crosses) and all TSSs (dots) show the same trend and that the increase in P-elements in the ±1000 bp region is not the result of including P-elements from neighboring TSSs. (B) Distance between neighboring TSSs in the genome, ranging from 1 to 5000 bp apart, showing that a large proportion (38.7%) of all TSSs in the genome are less than 2 Kb apart. The peak in distances between 1500 and 2000 bp refers to the tandem repeated histone genes.

Table 3.1 Distribution of P-element insertions in promoter regions.

We note that the number of P-element insertion sites for all TSSs is greater than the total number of insertion sites in the ±1 Kb around each TSS because of redundancy in counts from neighboring TSSs. This artifact is eliminated by restricting our analysis to TSSs that are greater than 2 Kb from any other TSS.

Dataset	Strand	-1000 to TSS	TSS to +1000	-1000 to +1000
# Insertion Sites in all TSS	Positive	4182	2340	6522
	Negative	4129	2406	6535
	Total	8311	4746	13057
# Insertion Sites in TSS >2 Kb apart	Positive	2464	1172	3636
	Negative	2426	1281	3707
	Total	4890	2453	7343

3.4.4 P-elements prefer to insert upstream of the TSS in nucleosome free regions.

We next investigated the fine-scale distribution of P-element insertions in promoter regions by plotting the P-element insertions around the TSS. We observed that the majority of P-element insertions in promoter regions occur between -190 and +80 bp around the TSS (Figure 3.4A). We also observed a strong preference for P-elements to insert upstream of the TSS (Figure 3.4A; Table 3.1). This result contrasts with previous analyses by Bellen et al. (Bellen, Levis et al. 2004) who reported that "P-elements strongly tend to insert within 100 bp symmetrically around the transcription start site." The bias to insert upstream of the TSS is statistically significant (Binomial Test: $P<2.2x10^{-16}$). The upstream bias is also observed when we restrict our analysis to genes whose TSSs are more then greater 2Kb away from each other (Binomial Test: $P<2.2x10^{-16}$), and thus this effect is not an artifact of insertions from nearby promoters. Furthermore, we observe a bias towards 5' insertion both inside (Binomial Tests: $P<2.2x10^{-16}$) and outside (Binomial Tests: $P<2.2x10^{-16}$) the main peak from -190 and +80 bp around the TSS, suggesting that cause of the upstream bias is not solely related to the factors that promote insertion in the -190 and +80 bp window.

The P-element tendency to insert preferentially upstream of the TSS may be linked with the nucleosome free region that has been shown to occur in the vicinity of the TSS in *D. melanogaster* (Mavrich, Jiang et al. 2008). In bulk nucleosomes, Mavrich et al. (2008) identified a nucleosome free region from -180 to +135 that could explain the peak in P-

element insertions we observe from -190 to +80. Using data on H2AZ containing nucleosomes extracted from embryos (Mavrich, Jiang et al. 2008), we calculated the overlap between P-element insertions and nucleosome-bound regions across the genome and specifically in promoters. Nucleosome-bound regions cover 30.99% of the entire *D. melanogaster* genome sequence (Table 3.2), with higher nucleosome coverage in promoters (37.90% in the ±1 Kb around TSS) relative to non-promoter regions (29.52% outside the ±1 Kb around TSS). Using the observed nucleosome coverage as the proportion of P-element insertions that would be expected to insert in nucleosome bound regions, we observe an apparent preference in P-element insertions into nucleosome bound DNA genome-wide. However this is an artifact of promoters having a higher proportion of nucleosome bound DNA and P-elements preferring to insert into promoter regions. When promoter and non-promoter regions are analyzed separately, we find evidence that the P-element avoids nucleosome-bound DNA in promoter regions (Binomial test; $P=1.98x10^{-14}$) but not outside promoter regions (Binomial test; $P=5.23x10^{-1}$). Within promoter regions there is a clear negative correlation (rho = -0.535 for Spearman's correlation test $P=3.347x10^{-16}$) between the location of P-element insertions and nucleosome density (Figure 3.4A vs Figure 3.4B; Figure 3.5).

Table 3.2 Transposon insertions in nucleosome-bound regions.

Percent of the entire genome, promoter regions and non-promoter regions covered by nucleosome coverage is shown in parentheses and reflects the expected proportion of transposons insertions that should randomly insert into nucleosome bound regions.

Transposable element	Genome (30.99)	Promoter (37.91)	Non-promoter (29.52)
P-element	32.68	33.90	29.99
piggyBac	20.57	26.71	16.49
Minos	22.04	23.76	21.47

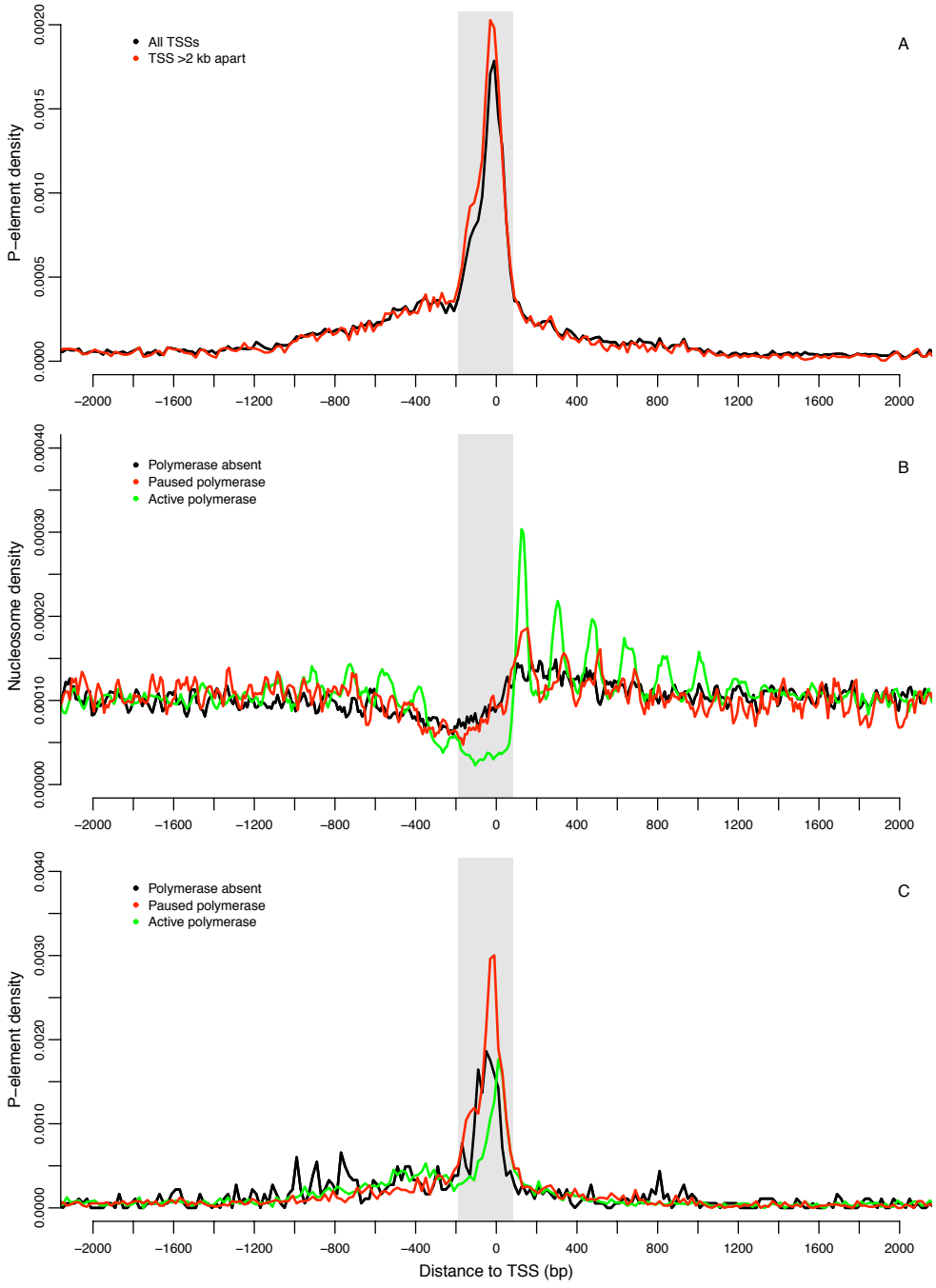


Figure 3.4 Distribution of P-element insertions and nucleosomes around the TSS

(A) P-element insertion density in 20 bp windows for all TSSs (black line) and for TSS more then 2 Kb away from another TSS (red line). The grey shaded area indicates the -190 to +80 region that has the highest P-element density. (B) Non-overlapping nucleosome density in 10 bp windows according to RNA polymerase status using data from (Zeitlinger, Stark et al. 2007). RNA polymerase absent (black line), RNA polymerase paused (red line) and RNA polymerase active (green line). (C) P-element density in 20 bp windows according to RNA polymerase state as in panel B. It is possible to see a shift in P-element insertion to occur more downstream of the TSS for both active and paused RNA polymerase. This effect is more pronounced in promoters with active RNA polymerase that also have a lower nucleosome density in these regions as shown in panel B.

3.4.5 Nucleosome avoidance is not specific to P-elements.

To address whether nucleosome avoidance is a specific property of the P-element that is a principal factor in explaining promoter targeting, we measured the frequency of insertions in nucleosome-bound regions for *piggyBac* and *Minos*, two transposons that do not show promoter targeting like the P-element. Both *piggyBac* and *Minos* show a stronger avoidance of nucleosomes than the P-element, both genome-wide and specifically in promoter regions (Binomial tests; $P<2.2x10^{-16}$and $1.56x10^{-11}$ respectively). Both *piggyBac* and *Minos* also presented a strong avoidance of nucleosomes outside the promoter regions (Binomial tests; $P<2.2x10^{-16}$). These results indicate that nucleosome avoidance may be a more general feature of DNA-based transposons, which is not specific to the P-element and its unusual target preference for promoter regions (Gangadharan, Mularoni et al. 2010). Therefore, we investigated additional factors associated with core promoter architecture that might be responsible for P-element promoter targeting.

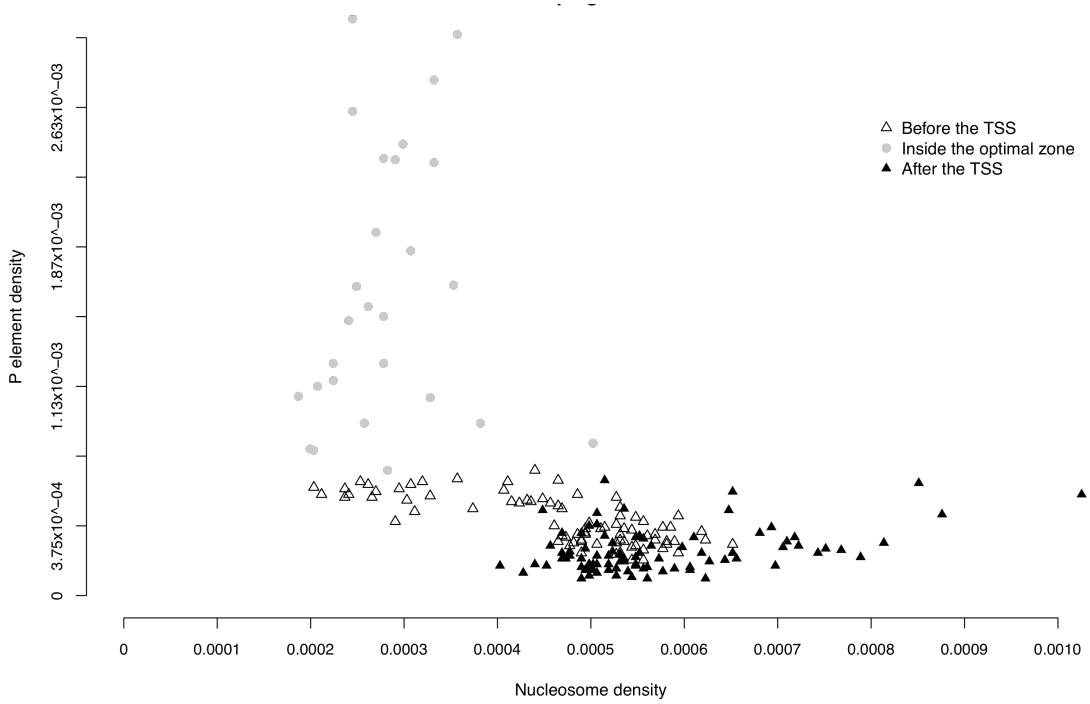


Figure 3.5 P-element density versus nucleosome density

P-element insertion density from ±1000 bp of the TSS in 10 bp windows plotted against non-overlapping nucleosomes in the same window for regions upstream of the TSS, in the peak of P-element insertion, and downstream of the TSS. It is possible to see the negative correlation between P-element and nucleosome density with the optimal zone for insertions presenting the lowest nucleosome density (below 0.04%) and the area after the TSS showing the highest nucleosome densities. The higher density of nucleosomes downstream of the TSS could explain the bias in P-element insertions upstream of the TSS.

3.4.6 RNA Polymerase pausing impacts P-element promoter targeting and insertion site location

The P-element has previously been reported to preferentially insert into Hsp genes (Lerman, Michalak et al. 2003; Shilova, Garbuz et al. 2006; Walser, Chen et al. 2006), which are classical examples of genes with paused RNA polymerase (reviewed in Fuda, Ardehali et al. 2009). To test if the P-element distribution is associated with genes that have different RNA polymerase status, we cross-referenced data on RNA polymerase pausing from Zeitlinger et al. (2007) and Muse et al. (2007) with our TSS and P-element data. For this analysis and those that follow, we performed preliminary investigations of the univariate effects of genomic factors on P-element occupancy using χ^2 tests of association with different datasets, which are reported in Table 3.3. We then chose Zeitlinger *et al.* (2007) data set with the DPE and TATA motifs from

JASPAR, with scores bellow -4.2 and -5.6 at positions +25 and -50 to -15 respectively, and the DRE motif from Ohler, with scores bellow -8 from position -120 to +20, to test for associations between the presence of P-elements and a particular promoter feature region using a multivariate generalized linear model (GLM) with binomial errors that included all factors, simplified to include only significant factors and interactions (Table 3.4). This analysis showed that presence of a P-element in a promoter was positively associated with both paused and active RNA polymerase at a TSS for both the Zeitlinger *et al.* (2007) and Muse *et al.* (2007) datasets (Table 3.3). Furthermore, these results remain significant when RNA polymerase status is considered jointly with other features of promoters (Table 3.4)

Table 3.3 χ^2 test test for individual genomic factors with P-element insertion in promoter regions.

Features	Enrichment	P value	#TSS with feature
Polymerase absent (Zeitlinger)	0.1344	< 2.2E-16	2577
Active Polymerase (Zeitlinger)	1.656	< 2.2E-16	2886
Paused Polymerase (Zeitlinger)	6.308	< 2.2E-16	1239
Polymerase absent (Muse)	0.1407	< 2.2E-16	4619
Active Polymerase (Muse)	3.312	< 2.2E-16	3824
Paused Polymerase (Muse)	9.877	< 2.2E-16	867
TATA (JASPAR)	0.5956	< 2.2E-16	2525
TATA (Ohler)	0.5668	< 2.2E-16	2638
TATA (Gershenzon)	0.5382	< 2.2E-16	2578
TATA (Zhu first analysis)	0.3123	< 2.2E-16	565
TATA (Zhu combined)	0.3249	< 2.2E-16	592
TATA (Zhu Patser)	0.2919	< 2.2E-16	673
DPE (JASPAR)	1.45	3.57E-06	659
DPE (Ohler)	1.619	1.55E-05	334
DPE (Gershenzon)	1.324	2.16E-04	744
DPE (Zhu first analysis)	1.343	1.59E-01	105
DPE (Zhu combined)	1.343	1.59E-01	105
DPE (Zhu Patser)	1.621	1.06E-06	427
TRF2	3.973	< 2.2E-16	1941
TBP	2.573	< 2.2E-16	540
DRE (Ohler)	1.756	< 2.2E-16	2805

DRE (Zhu first analysis)	1.874	< 2.2E-16	1225
DRE (Zhu combined)	1.773	1.11E-15	839
DRE (Zhu Patser)	1.803	< 2.2E-16	1855
DSP1	6.457	< 2.2E-16	2237
GAF	6.922	< 2.2E-16	3091
H3K4me3	5.804	< 2.2E-16	8271
H3K27me3	0.5423	1.26E-06	311
PC	0.9461	5.60E-01	550
PH	0.854	3.03E-01	208
PHO	4.386	< 2.2E-16	3811
PHOL	4.737	< 2.2E-16	4935
TRX.C	1.318	2.01E-01	100
TRX.N	4.872	< 2.2E-16	6566

Table 3.4 The final GLM with polymerase data from Zeitlinger et al. (2007) with the DPE and TATA motifs from JASPAR and DRE motif from Ohler (2002).

This model was simplified by removing non-significant variables and interactions from the full GLM.

Genomic Factor	Model Coefficient	P-value
TATA	-0.318	2.63E-02
DPE	0.287	5.79E-02
DRE	0.255	4.26E-04
TRF2	1.249	3.30E-05
Recruiter PcG	1.834	< 2.2E-16
Other PcG	-0.779	2.56E-07
trxG	1.093	< 2.2E-16
Active Polymerase	0.455	1.05E-06
Paused polymerase	1.763	< 2.2E-16
DPE : TRF2	-0.972	2.48E-02
TATA: Recruiter PcG	1.611	9.95E-04
Recruiter PcG: trxG	0.657	2.54E-03
TRF2 : trxG	-0.767	1.45E-02
Recruiter PcG: trxG	-1.017	2.20E-07
TATA: Recruiter PcG : trxG	-1.384	4.30E-03

Given our results that nucleosome positioning influences P-element insertion, together with previous work by (Mavrich, Jiang et al. 2008) who showed that nucleosome positioning is affected by polymerase pausing, we asked if the distribution of P-element insertion sites was influenced by polymerase pausing. As shown in Figure 3.4C, P-element insertion locations vary according to RNA polymerase status (no polymerase present, active polymerase, paused polymerase), with significant differences in the distribution of P-element insertions in the peak from -190- to +80 (Table 3.5 and Table 3.6). For both the Zeitlinger *et al.* (2007) (Figure 3.4B) and Muse *et al.* (2007) (Figure 3.6A) datasets, we observe that P-element insertions are most biased to the 5' when no RNA polymerase is present. When RNA polymerase is present, P-elements tend to insert further downstream, with a greater shift downstream in promoters with actively transcribing polymerase relative to a paused polymerase (Table 3.5 and Table 3.6). Shifts in the location of P-element insertion correlate with the depletion of nucleosomes in the -190 to +80 window, such that when the nucleosome coverage in this region decreases the number of insertions goes up and vice versa (Figure 3.4C, Figure 3.6B). These results support the conclusions that P-elements tend to avoid nucleosomes (see above) and that the P-element transposition complex can detect important features of core promoter architecture such as the configuration of the RNA polymerase.

Table 3.5 P-element insertion patterns in TSSs of promoters with different RNA polymerase status using data from Zeitlinger *et al.* (2007).

P-values are for nonparametric tests for differences in the central tendency (Wilcoxon Test) and distribution (Kolmogorov-Smirnov test) of P-element insertion locations in the window from -190 to +80 around the TSS.

RNA polymerase status	Wilcox test	Kolgomorov test	Difference of median location (bp)
Absent vs. paused	1.49E-02	8.73E-03	-17
Absent vs. present	5.04E-14	1.04E-14	-44
Paused vs. present	<2.2E-16	<2.2E-16	-28

Table 3.6 P-element insertion patterns in TSSs of promoters with different RNA polymerase status using data from Muse *et al.* (2007).

P-values are for nonparametric tests for differences in the central tendency (Wilcoxon Test) and distribution (Kolmogorov-Smirnov test) of P-element insertion locations in the window from -190 to +80 around the TSS.

RNA polymerase status	Wilcox test	Kolgomorov test	Difference of median location (bp)
Absent vs. paused	9.54E-13	1.11E-09	-20
Absent vs. present	< 2.2E-16	< 2.2E-16	-37
Paused vs. present	6.25E-11	9.90E-12	-17

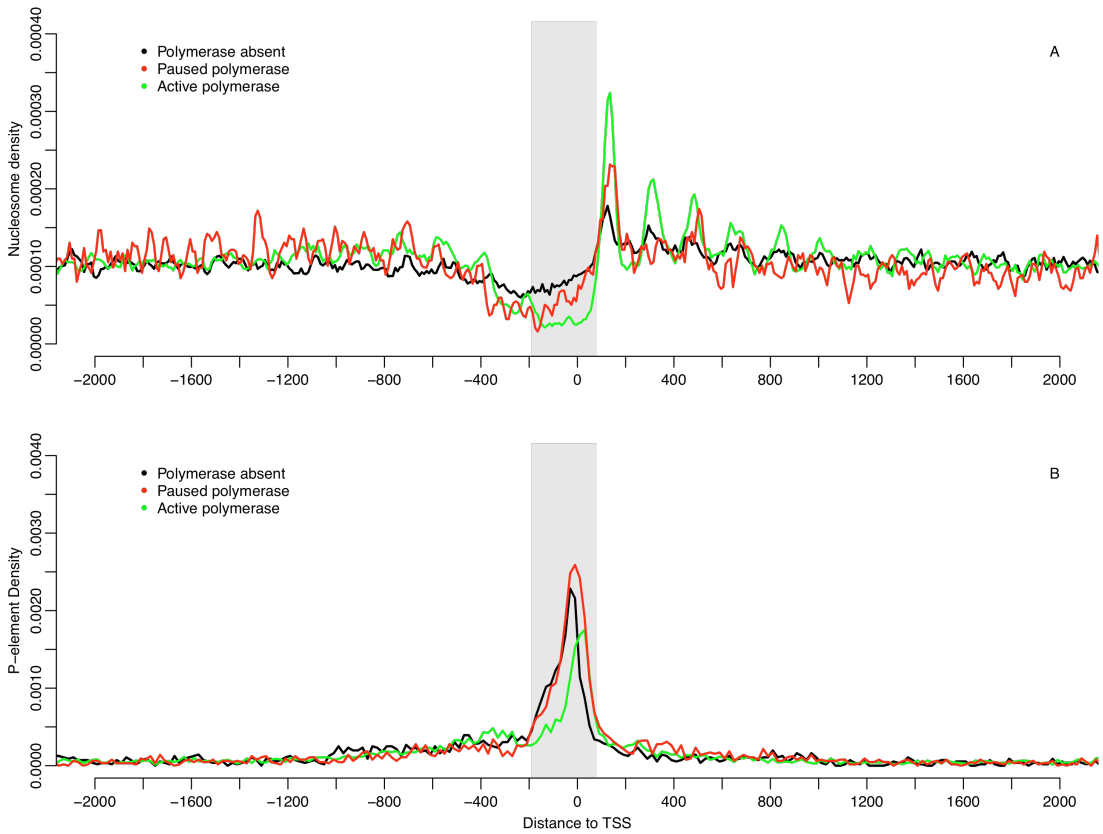


Figure 3.6 P-element nucleosome avoidance in the polymerase dataset from Muse et al. (2007).

(A) Non-overlapping nucleosome density in 10 bp breaks for the data from Muse et al. (2007) according to the polymerase state; absent (black line), paused (red line) and active (green line). (B) P-element density in 20 bp windows according to polymerase state has above. The grey shade indicates the same region (-190 to +80) in both graphs. Although the data from Zeitlinger et al. (2007) overlaps the other datasets it does not coincide with the time of P-element transposition (from 8 to 24 hours embryonic development Engels and Preston 1979). The analysis of the dataset from Muse et al. (2007) confirms the same relation between RNA polymerase and P-element insertions.

3.4.7 P-elements prefer TATA-less promoters

Since the presence of a paused polymerase has been show to correlate with specific core promoter motifs (Hendrix, Hong et al. 2008), we further investigated whether core promoter motif architecture differed in promoters with and without P-elements. As a first approach that did not require motif prediction, we computed the base composition of promoter regions with (Figure 3.7A) and without (Figure 3.7B) P-elements. We quantified the difference between these two classes of promoters, by calculating a χ^2 statistic for each position in the core promoter region that is known to contain the major core promoter motifs (±60 bp from the TSS). This analysis revealed that there were

significant differences in base composition at many positions between promoters with and without P-elements (Figure 3.7C, black line). Interestingly, differences are not distributed randomly across the entire core promoter but are instead concentrated in specific positions ranging from -30 to -20 and from -1 to +31, regions which contain the TATA box and INR/DPE motifs, respectively. The positions that showed largest differences in base composition between promoters with and without P-element were positions -28 (containing TATA) and -1 (containing INR). To address the possibility that these differences in base composition could occur by chance, we generated 1,000 randomized datasets of TSSs of the same size as those with (n=6,005) and without (n=8,224) P-elements. We then calculated a χ^2 statistic for each randomized dataset and plotted the upper 97.5% confidence interval of χ^2 values (Figure 3.7C, grey line). This analysis revealed that all positions with χ^2 values greater than 20 are likely to represent statistically significant differences in base composition between promoter classes. Thus, we conclude that there are clear differences in the base composition between promoters of genes with and without P-elements, which are localized to specific sub-regions of the core promoter that contain known promoter motifs.

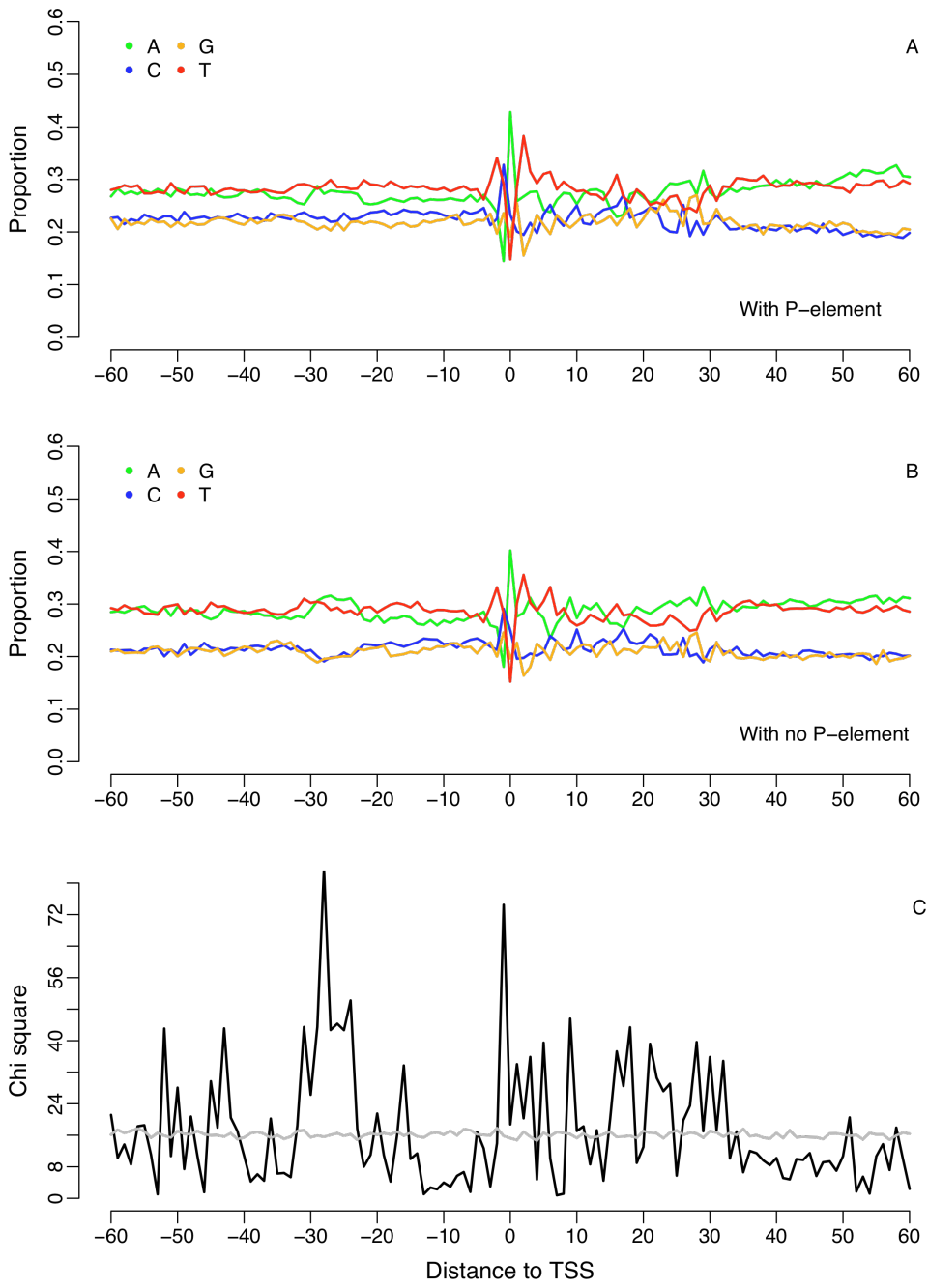


Figure 3.7 Base composition and χ^2 test for P-element targeted and non-targeted TSSs

(A) Nucleotide compostion in a -60 to +60 bp window for the 6,005 TSSs that are associated with a P-element. (B) Nucleotide density for the 8,224 TSSs that are not associated with a P-element. C) χ^2 test test for the difference between nucleotide composition for P-element and non P-element associated TSSs at the bp level (black line) and the top 97.5 percentile from 1,000 simulations of a random selection of TSSs the same size has in A and B. The χ^2 test value for the top 97.5 percentile has a mean of 16.156 and all the χ^2 test values below it have a P-value lower than 7.6E-4. The broadest and highest peak region of differences reflects the location of the TATA box and INR/DPE regions.

To test directly whether the P-element targets promoters that contain specific motifs, we developed a method to annotate and classify promoters based on their predicted promoter motif composition using position weight matrices (PWMs). We restricted this analysis to the best-studied core promoter motifs, TATA and DPE, because PWM-based promoter motif prediction is an inexact process (dependent on the choice of PWM, search window and match score cut-off), and we could only optimize our promoter motif annotation strategy using manually-curated promoter motifs for TATA and DPE from the DCPD (Kutach and Kadonaga 2000) (see above). Using PWMs from the JASPAR database with optimized windows and PWM match score cut-offs, we predicted the presence or absence of TATA and DPE motifs for all 14,229 non-redundant TSSs in our data set. To avoid bias in the choice of PWM and to assess how robust our results are to independent promoter motif annotation methods, we also analyzed a previously published set of motif annotations from Zhu *et al.* (2009) for 12,588 TSSs of their 16,513 TSSs that mapped exactly to our to TSSs.

This analysis revealed that the presence of a P-element in a promoter region is negatively correlated with the presence of a TATA motif and positively correlated with the presence of a DPE motif (Table 3.3; Table 3.4). This result is observed for all motif annotation sets analyzed (Table 3.3) and is consistent with the analysis of base composition above, which clearly shows a decrease in the AT composition in the region where the TATA motif should be in P-element containing promoters (Figure 3.7A vs Figure 3.7B). Together these results indicate that the P-element prefers to insert into TATA-less promoters. However, it is important to note that the strength of the P-element association with core promoter motifs is relatively weak and that the P-element can be found in all four major types of promoter (TATA+/DPE+, TATA-/DPE+, TATA+/DPE- and TATA-/DPE-).

3.4.8 P-elements prefer TFR2-bound, DRE-containing promoters

TATA-containing promoters typically recruit RNA polymerase directly through TBP, while TATA-less promoters can either recruit RNA polymerase through TBP associated factors (TAFs) or through the TBP-related factor (TRF2) (Isogai, Keles et al. 2007). Isogai *et al.* (2007) have performed whole genome ChIP-chip experiments in *Drosophila* S2 cells with both TBP and TFR2, allowing us to investigate whether P-

element insertions are positively associated with promoter regions bound by either of these GTFs. Consistent with the motif prediction analyses above, we find that the P-element is most strongly associated with TRF2-bound promoters (Table 3.3; Table 3.4), which have a low frequency of TATA motifs (Isogai, Keles et al. 2007). We also observed a weak positive correlation between P-elements and TBP-bound promoters (Table 3.3), however this association did not remain significant in the joint GLM. This result may indicate a spurious correlation between TBP and P-element insertion, since many promoters are bound by both TBP and TRF2 in S2 cells (n=243 for TRF2 and TBP bound promoters out of 1,941 and 540 regions respectively) (Isogai, Keles et al. 2007).

TRF2 does not bind DNA directly (Rabenstein, Zhou et al. 1999), but is thought to bind DNA as part of a complex through DNA-replication element (DRE) binding factor DREF (Hochheimer, Zhou et al. 2002). As a consequence, TRF2 bound promoters are known to have a high frequency of the DRE motif (Isogai, Keles et al. 2007). To support the preferential association of P-elements with TRF2-bound promoters we tested if P-elements are more often found in promoters that contain a DRE motif by predicting DRE motifs in our TSS dataset using a PWM-based strategy. Unlike the analysis of TATA and DPE motifs above, we were not able to optimize our own motif prediction method for DRE on a curated training dataset. Thus, we used previous search criteria reported in Zhu and Halfon (2009) to generate our own DRE motif predictions, as well as using motif annotations from Zhu and Halfon (2009) for the set of their TSSs that mapped exactly to our TSSs. Regardless of motif annotation method, we find that the presence of P-elements is positively correlated with the presence of a DRE motif in promoter regions (Table 3.3; Table 3.4) as would be expected if P-elements insert into TRF2 bound promoter regions.

3.4.9 Promoters containing H3K4me3 modified histones and Polycomb recruiter proteins are targeted by the P-element

In addition to DREF, TRF2 is associated *in vitro* with subunits of the nucleosome-remodeling factor (NURF) complex (Hochheimer, Zhou et al. 2002). NURF is part of the *trithorax* group (trxG) of proteins (Schuettengruber, Chourrout et al. 2007) and is associated with trimethylation of Histone H3 on lysine 4 (H3K4me3) (Wysocka, Swigut

et al. 2006), an epigenetic marker of active chromatin. Functional links between TRF2, DREF and the NURF complex in active promoter regions are implicated in the recent genome-wide ChIP-chip analysis of (Schuettengruber, Ganapathi et al. 2009), who found that the DRE motif was increased in the promoter regions of genes with high levels of H3K4me3 located at their TSS. In addition, members of the *Polycomb* group (PcG) of repressor proteins that counteract trxG activation colocalize to TSS regions and have been found in stable complexes with members of the general transcriptional machinery.

Therefore, we investigated possible connections between P-element promoter targeting and trxG and PcG factors and associated epigenetic modifications by mapping the ChIP-chip data from Schuettengruber et al. (2009) to our TSS dataset. This dataset includes genome-wide location data for H3K4me3- and H3K27me3-modified nucleosomes, Pleohomeotic (PHO), PHO like (PHOL), GAGA factor (GAF), Dorsal switch protein (DSP1), Polycomb (PC), Polyhomeotic (PH), and two parts of the TRX protein (C- and N-terminal regions). This analysis revealed significant positive associations between the P-element and factors associated with active chromatin marks such as H3K4me3 and the N-terminal part of TRX, as well as PcG proteins referred to as the "recruiters" (Schuettengruber, Ganapathi et al. 2009), including DSP1, GAF, PHO and PHOL proteins (Table 3.3). We found significant negative associations between P-element insertion in promoter regions and the repressive chromatin mark H3K27me3. We also found non-significant negative associations between P-element insertion and the binding of both PC and PH, which are members of the PRC1 complex that recognizes the H3K27me3 mark and exerts PC-mediated silencing. To model the biology of these complexes more realistically and bypass the correlated effects of factors that bind overlapping regions in the genome, we reduced these 10 ChIP-chip datasets into the 3 functional groups defined by Schuettengruber et al. (2009) – recruiter PcG, non-recruiter PcG, and trxG – and classified TSSs as belonging to these groups if any of the factors in each group were present in a TSS. As shown in Table 3.4, recruiter PcG and trxG factors are positively associated with the P-element in TSSs while the non-recruiter PcG proteins are negatively associated, even when considered together with other genomic factors. The presence of recruiter PcG and trxG groups is highly predictive of P-element insertion**,** with only 619 out of 6,005 promoter regions (10.3%) that are targeted by P-elements not having either a PcG recruiter or trxG binding region

(P<2.2E-16 for the Chi-test). P-element association with PcG/trxG proteins does not seem to be solely due to proximity to TSS since 69.9% of all P-element insertions genome wide (8,567/12,267) are in a ±1 bp window from a PcG/trxG protein binding region. The association with PcG recruiter or trxG binding also seems to be exclusive to the P-element, since both *piggyBac* and *Minos* show significantly lower proportions of insertion events associated with PcG/trxG group binding regions, 34.1% (3,972/11,633) and 4.7% (96/2,052) respectively (Binomial Tests: P-Values <2.2e-16).

3.4.10 Genes expressed in the female germline and S2 cells are susceptible to P-element insertions

Thus far we have considered how genomic factors relating to the sequence or chromatin state affect P-element insertion in a promoter region, since these are the factors that could causally influence P-element insertion into the genome. However, particular genomic factors like motif composition have previously been shown to be associated with tissue-specific expression (FitzGerald, Sturgill et al. 2006), and since P-element insertion is associated with specific promoter features, we also investigated whether P-element insertion might also be associated with tissue specific expression. We analyzed expression separately and did not include expression into our joint GLM, because gene expression, like P-element insertion, is an output of the different genomic factors that are inputs to the process of transcriptional regulation, and therefore cannot be a causal factor determining P-element promoter targeting.

To test for associations between P-element insertion and gene expression, we mapped tissue-specific expression data from FlyAtlas (Chintapalli, Wang et al. 2007) to our dataset of TSSs. FlyAtlas includes gene expression data from 26 different tissues. All tissues showed a strong positive association between expression and P-element insertion (Table 3.7**)** with the exception of head (no significant association) and testis (significant negative association). The strongest positive association for P-element insertion is for genes expressed in ovaries and S2 cells. Similar correlations can be seen between binding of PcG recruiter/trxG proteins and gene expression, with testis and head being the only two tissues that do not show a positive association (Table 3.8). In other words, when a tissue was positively associated with P-element insertion it was also positively associated with the PcG recruiter/trxG binding and the converse was also true.

Table 3.7 χ^2 test for association between tissue specific gene expression and P-element insertion into promoter regions.

Tissue	Enrichment	Correlation P value	#TSS with feature
Brain	1.226	< 2.2E-16	1915
Head	1.05	4.48E-02	1601
Crop	1.315	< 2.2E-16	1545
Midgut	1.13	4.26E-06	1311
Hindgut	1.258	< 2.2E-16	1562
Tubule	1.198	1.43E-15	1591
Ovary	1.349	< 2.2E-16	2472
Testis	0.7138	< 2.2E-16	1477
Accessory gland	1.223	6.96E-16	1358
Larval tubules	1.258	< 2.2E-16	1473
Larval fat body	1.219	5.62E-11	996
Thoracic ganglion	1.222	< 2.2E-16	1790
Carcass	1.111	5.99E-05	1363
Salivary glands	1.305	< 2.2E-16	1514
Larval salivary glands	1.334	< 2.2E-16	1495
Larval midgut	1.141	2.58E-06	1199
Larval hindgut	1.27	< 2.2E-16	1512
Virgin spermatheca	1.19	1.12E-11	1331
Mated spermatheca	1.239	< 2.2E-16	1316
Larval CNS	1.269	< 2.2E-16	2304
Adult fat body	1.201	2.89E-12	1268
Larval carcass	1.243	< 2.2E-16	1229
Eye	1.195	8.62E-15	1560
Heart	1.213	3.17E-14	1325
Larval trachea	1.356	< 2.2E-16	1577
Drosophila S2 cells	1.375	< 2.2E-16	1985

Table 3.8 χ^2 test for association between tissue specific gene expression and PcG recruiter or trxG binding.

Tissue	Enrichment	Correlation P value	#TSS with feature
Brain	1.215	< 2.2E-16	1915
Head	0.9451	3.50E-04	1601
Crop	1.187	< 2.2E-16	1545
Midgut	1.057	1.24E-03	1311
Hindgut	1.123	2.79E-15	1562
Tubule	1.146	< 2.2E-16	1591
Ovary	1.443	< 2.2E-16	2472
Testis	0.7709	< 2.2E-16	1477
Accessory gland	1.267	< 2.2E-16	1358
Larval tubules	1.231	< 2.2E-16	1473
Larval fat body	1.155	5.84E-14	996
Thoracic ganglion	1.212	< 2.2E-16	1790
Carcass	1.001	9.84E-01	1363
Salivary glands	1.251	< 2.2E-16	1514
Larval salivary glands	1.337	< 2.2E-16	1495
Larval midgut	1.049	8.83E-03	1199
Larval hindgut	1.132	< 2.2E-16	1512
Virgin spermatheca	1.137	2.28E-15	1331
Mated spermatheca	1.167	< 2.2E-16	1316
Larval CNS	1.304	< 2.2E-16	2304
Adult fat body	1.129	4.00E-13	1268
Larval carcass	1.105	7.52E-09	1229
Eye	1.158	< 2.2E-16	1560
Heart	1.141	4.13E-16	1325
Larval trachea	1.261	< 2.2E-16	1577
Drosophila S2 cells	1.354	< 2.2E-16	1985

3.5 Discussion

The use of TEs in the creation of mutations and in the study of expression patterns has been a fundamental part in understanding the function and organization of genomic sequences of *D. melanogaster*. One of the most widely used TE in these studies is the P-element, whose main drawback for genome-wide analysis has been its specificity to target promoter regions of a subset of genes. In this study we show that the nonrandom insertion of P-elements into promoters can be linked to aspects of core promoter architecture including nucleosome positioning, motif composition, binding of general transcription factors and chromatin modifying factors of the *Polycomb* and *trithorax* groups.

3.5.1 Nucleosome avoidance shapes distribution of P-element insertion in promoter regions but does not explain promoter targeting.

Our analysis of the fine scale spatial pattern of P-element insertion revealed that P-elements preferentially insert upstream of the TSS. This is true both in the main peak of P-element insertion from -190 to +80 and in the broader region of increased P-element insertion that extends to ±1000 bp around the TSS. The same pattern is observed in TSSs that are >2 Kb from any other, revealing that this preference is unlikely to be an artifact caused by insertions in neighboring promoter regions. One trivial explanation for insertion upstream of the TSS could be if mRNAs had incompletely annotated 5' UTRs and the TSS would be incorrectly placed to the 3' of its true location. However the annotation procedures used by FlyBase appear to have the opposite bias, with the most 5' EST/cDNA being used to define the mRNA start site (Rach, Yuan et al. 2009). Moreover, the curated DCPD dataset (Kutach and Kadonaga 2000) also shows the same pattern of insertion (Figure 3.8). We interpret the preference to insert upstream of the TSS to be related with the nucleosome free region that exists in the -180 to +135 region of *D. melanogaster* promoter regions (Mavrich, Jiang et al. 2008). In support of this conclusion, we find that the P-element avoids insertion into nucleosome bound regions in promoter regions. Moreover, we see changes in the fine scale distribution of P-element insertion that correlate with the presence of a paused RNA polymerase, which is associated with shifts in the distribution of nucleosomes in promoter regions.

However we also find that both *piggyBac* and *Minos* exhibit nucleosome avoidance, neither of which target promoters, and thus we conclude that nucleosome avoidance cannot be the main causal factor in determining P-element promoter targeting in *Drosophila*.

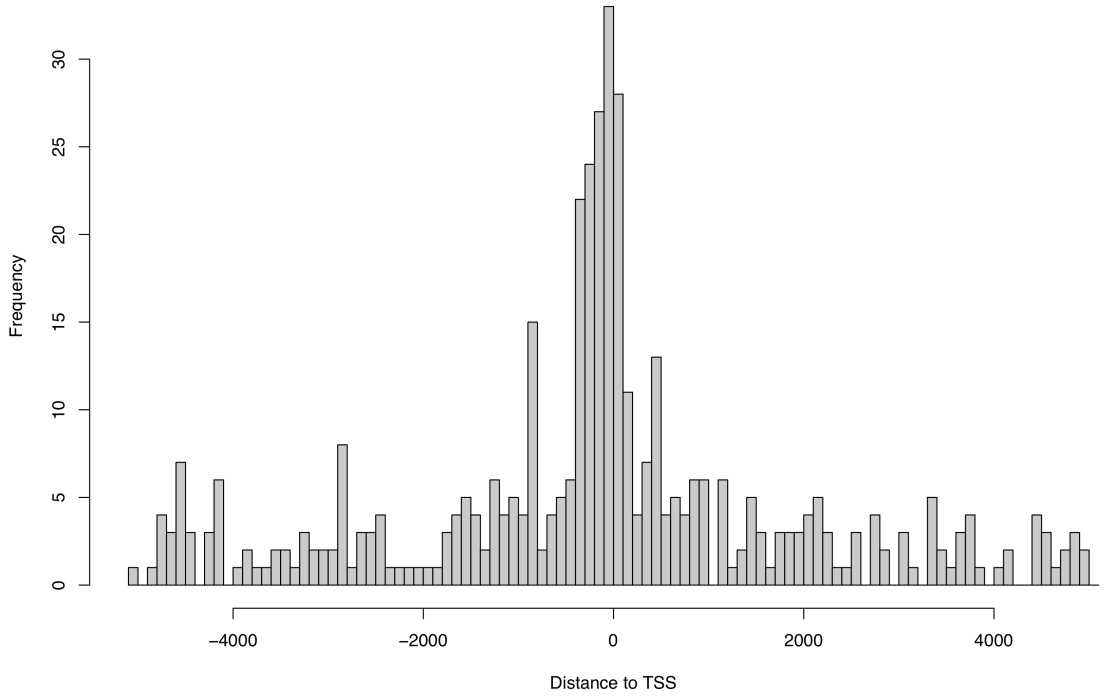


Figure 3.8 P-element distance to the annotated TSSs from DCPD

Frequency histogram of the P-element non-redundant insertion sites close to the 197 annotated TSSs that were close to our set of TSSs. It is possible to see the same P-element target site preference has in figure 3.4A with a left skewed preference for TSSs.

3.5.2 RNA polymerase activity affects P-element insertion sites close to genes

In addition to nucleosome positioning, P-element insertion in promoter regions appears to be influenced by the presence and activity of RNA polymerase. The P-element preference for promoters with active or paused RNA polymerase could be explained by the fact that the presence of RNA polymerase at the TSS requires chromatin to be decondensed, allowing the P-element transposase access to the DNA. When RNA polymerase is paused, promoter regions are apparently more open and accessible for P-element insertion, resulting in a stronger positive association than for active polymerase. When RNA polymerase is active there is a downstream shift in the position of the +1 nucleosome and consequently a shift in P-element insertion sites. Together, these results

indicate that the P-element transposase complex can detect subtle aspects of promoter architecture that are dependent on RNA polymerase activity.

3.5.3 P-element promoter motif preferences

We also find associations between P-element insertion and the presence of specific core promoter motifs. The observed negative association between P-element insertion and TATA motifs could be related to the P-element preference for a CG rich motif (Linheiro and Bergman 2008) or with the fact that TATA is enriched in TSSs of genes that use a unique promoter (Zhu and Halfon 2009). Both of these factors could contribute to the reduced likelihood of a target site for P-element insertion. In contrast, the DPE motif presented a positive association with P-element insertion. This promoter motif is CG rich (Kadonaga 2002; Smale and Kadonaga 2003; FitzGerald, Sturgill et al. 2006), like the P-element target site motif, and has also been associated with promoters that have a paused RNA polymerase (Hendrix, Hong et al. 2008), factors that increase the likelihood of a target site and an easier access to the DNA. The promoter motif that showed the strongest correlation with the P-element was the DRE motif. This motif is associated with gene expression in the female germline (FitzGerald, Sturgill et al. 2006) and with PcG and TRF2 bound promoters (FitzGerald, Sturgill et al. 2006; Isogai, Keles et al. 2007; Schuettengruber, Ganapathi et al. 2009; Zhu and Halfon 2009), which are two other factors we find that increase the likelihood of P-element insertion in promoter regions.

3.5.4 General transcription factor association with P-element insertions

Supporting a role for active transcription, we found that TRF2 binding showed a positive association with P-element insertion. The association with TRF2 was consistent with the fact that this transcription factor has been associated with TATA-less, DRE containing promoters (Isogai, Keles et al. 2007). The weak association of P-element insertion with TBP is somewhat surprising, but given the low number of TBP bound promoters and the overlap between TBP and TRF2 binding this may be a result of low statistical power. Alternatively, this result may reflect the preference of the P-element to avoid promoters containing TATA motifs, which are common in TBP bound promoters (Isogai, Keles et al. 2007). However TBP does not bind uniquely to TATA containing

promoters, since it can form an unstable complex with TATA-less promoters (Smale and Kadonaga 2003), and future studies may still detect an association between TBP and P-element insertion.

3.5.5 P-element association with the PcG TRX group proteins

One of the strongest genomic features that is associated with P-element insertion is the presence of recruiter PcG protein binding. This result makes sense because modification of chromatin is one of the first steps in promoter activation, and would therefore also be very important in P-element insertion. We think the association between PcG recruiter binding is fundamental to P-element insertion since there is an association between P-element insertion and PcG sites outside of promoter regions and other transposable elements used in *Drosophila* that do not show promoter targeting, *PiggyBac* and *Minos*, have much lower rates of insertion in recruiter PcG and trxG sites. The strong association of P-element insertion in promoters with the PcG is perhaps not surprising since it has been previously noticed that P-element "homing", that is the tendency that a P-element carrying part of a promoter region of a gene has to insert near that same gene, involved polycomb response elements and is dependent on PcG protein activity (Kassis, VanSickle et al. 1991; Kassis, Noll et al. 1992; Taillebourg and Dura 1999; Bender and Hudson 2000).

3.5.6 Gene expression and P-element

From first principles, it is expected that a TE should be active in the germline and not the soma, since only germline transposition events are inherited. Therefore genes expressed in the germline might be more likely to be targets of P-element insertion than those expressed in somatic tissues. Bownes (1990) has reported a preference for the P-element to insert into genes that are active during the male germline and Fontanillas, Hartl et al. (2007) have shown a positive correlation of germline expression on the frequency of non-P-element TE insertions into a gene. By analyzing the different expression datasets we were able to identify a strong negative correlation between the P-element and testis expressed genes and a positive association with ovaries expressed genes. Two factors that might relate to the negative correlation between P-element and testis expression might be the very different base composition pattern of testis-

expressed genes (Figure 3.9) and/or the negative correlation of testis expression with recruiter PcG proteins. The negative association between P-elements and male germline expression might also be related with a genetic imprinting effect since the P-element can only transpose in the cross between a P cytotype male and an M cytotype female (Rio 2002). The strong correlation between ovary expressed genes and P-elements may be linked with increased chance for expression of P-element transcripts in the egg, either to promote transposition or repression.

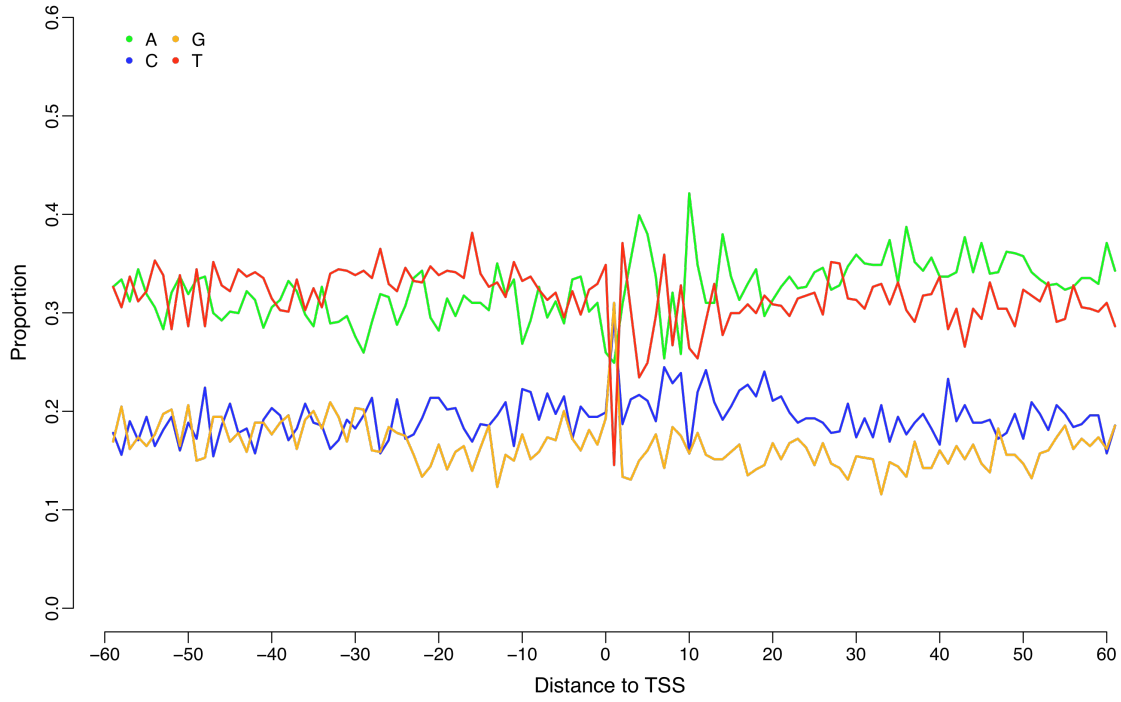


Figure 3.9 Base composition of testis specific promoters

Base composition for the 674 TSS whose transcripts are uniquely expressed in testis in a ± 60 bp window. Contrasted with base composition for all genes (Figure 3.7 A and B) there is no increase in the CG content close to the TSS of testis expressed genes, and have AT/CG levels close to the background level in *D. melanogaster*, 29 and 21% respectively.

3.5.7 P-element does not work alone

We have found evidence for several factors that affect the tendency for P-elements to specifically insert into promoter regions. These results may mean that the P-element does not work alone, requiring other host proteins to open or otherwise modify the chromatin at the insertion sites (Mullins, Rio et al. 1989). One such candidate group is the recruiter PcG proteins. Both PcG and trxG protein groups dissociate from the

genome in late G2 (reviewed in Ringrose and Paro 2007), the stage in the cell cycle when the P-element is also cut from the genome (Engels, Johnson-Schlitz et al. 1990; Weinert, Min et al. 2005). It is possible that if the PcG and trxG protein groups interact outside of the genome with the P-element, when attaching to the genome between anaphase and G1 (reviewed in Ringrose and Paro 2007) they would redirect the P-element to regions of the genome (e.g. the TSS of genes) that are associated with these protein groups. When analyzing the PcG and trxG proteins individually in a GLM the protein with the highest coefficient is GAF (Table 3.4). The GAF protein, like the P-element transposase (Rio 2002), has a zinc finger motif and binds to the consensus sequence GAGAG that can also be found at the P-element 5' end from position 85 to 90 overlapping the P-element TSS (O'Hare and Rubin 1983). The GAGAG motif is usually found at the 5' end of promoters from -140 to -61 from the TSS (Zhu and Halfon 2009) and GAF is also known to induce chromatin remodeling (reviewed in Schuettengruber, Ganapathi et al. 2009). From our analysis it seems that GAF may be a strong candidate for a host factor necessary for determining P-element insertion. In the future, it may be possible to test whether P-element promoter targeting is dependent on GAF or other PcG recruiter factors, by mobilizing the P-element in mutant backgrounds and observing if promoter targeting is disrupted.

4 Natural target site motif preferences of *D. melanogaster* transposable elements

4.1 Abstract

Transposable elements are mobile DNA sequences that are a source of mutations and target specific sites. The natural target preference of most TEs is unknown and is inferred after the insertion event occurred. Using genome resequencing data from 176 strains of *Drosophila melanogaster* gathered by the DGRP project we were able to identify 11,976 TE insertions in 8,033 new insertion sites that can be used to decode the natural target preference for DNA, LTR and non-LTR elements in this species. These insertions are not present in the reference strain and therefore represent recent insertion events and reveal the genomic context in to which they inserted. The insertions are non-uniformly distributed with some elements showing a greater degree of occupancy in the same insertion site. Both Illumina and 454 sequencing platforms showed consistent results in terms of target site duplication (TSD) and target site motif (TSM) discovery. TSMs typically extend the TSD and are palindromic for both DNA and LTR elements whose palindrome center varies according to the length of the TSD. Additionally, we found that TEs from the same subclass present similar TSDs and TSMs. Using the P-element as a benchmark, we show that there is overlap in target site preferences between artificial and natural insertion events. Our results demonstrate the utility of population genomics data for better understanding the targeting preferences of TEs in the wild.

4.2 Introduction

TEs are mobile DNA segments that occur within a host genome whose insertion and excision can cause disruption of genes and chromosomal rearrangement. They are considered a source of genetic variability and can be found in almost every organism from prokaryotes to eukaryotes (Biemont and Vieira 2006). According to their method of transposition TEs can be categorized in to two major classes: (i) those that transpose directly in to the host genome, *via* a DNA molecule (transposons), and (ii) those that transpose through an RNA intermediate (retrotransposons) (Craig 2002). Retrotransposons can be further subdivided into long terminal repeat (LTR) elements and non-LTR elements.

A characteristic mark of TE activity in the genome are TSDs, which occur upon TE insertion as a result of the staggered double strand breaks at the target site (Craig 2002). Both DNA and LTR elements insert in to the target site as a DNA-protein complex that causes a consistent length staggered cut that is characteristic of the TE family (Craig 2002). In the case of the DNA elements, it is a direct cut-and-paste process; in the case of the LTR elements, the TE is transcribed and reverse transcribed before inserting in to the target site. In contrast, transposition of non-LTR elements transposition leaves a variable length staggered break in the genome that leads to a distribution of TSD lengths for a given family (Eickbush and Malik 2002).

Understanding the nature of TSDs is important for several reasons. Since the TSD limits the extent of LTR and DNA elements in the genome, knowledge of TSDs can be used to further annotate the TE in the genome. For example, tools like LTRharvest (Ellinghaus, Kurtz et al. 2008; Fiston-Lavier, Carrigan et al. 2010) use the TSD among other characteristics to discover new TEs in the genome. The TSD can also be used to characterize a new family of either DNA or LTR elements since it is conserved throughout the subclasses (Bowen and McDonald 2001; Kapitonov and Jurka 2003). Finally, understanding of TSD properties can provide further insight into the general process of transposition for a family or class of TEs.

Traditionally, TSDs are discovered using methods that are based on analysis of sequences flanking TE insertions caused by *de novo* spontaneous mutations (O'Hare and

Rubin 1983; Fawcett, Lister et al. 1986; Mori, Benian et al. 1988; Viggiano, Caggese et al. 1997), TE insertions generated by artificial mutagenesis (Engels, Johnson-Schlitz et al. 1990; Tudor, Lobocka et al. 1992; Collins and Anderson 1994) and TE insertions found in genomic sequences (Bowen and McDonald 2001). Methods relying on spontaneous mutations or genomic sequences typically use only a few TE insertions to deduce the TSD and often provide data only for a single TE (Viggiano, Caggese et al. 1997) or a class of TE (Bowen and McDonald 2001). For methods that use artificial mutations, it is not usually possible to compare discovered TSDs to those from natural transposition events.

Here we develop a high-throughput method for identify new TSDs and TSMs based on *de novo* TE insertions using next-generation sequence data from whole genome shotgun resequencing projects. All that is required for our method is a reference genome and next generation sequencing reads long enough to include the start or end of a TE and its unique genomic flanking sequence. We apply our approach to *D. melanogaster,* a species that has a diverse range of TE families that have been previously characterized (Kaminker, Bergman et al. 2002). There are over 20 DNA families, 60 LTR families and 40 non-LTR families currently documented for this species, representative of 3 LTR subclasses and 8 DNA and non-LTR subclasses. The abundance of different families from different subclasses in *D. melanogaster* makes this species a good candidate to develop our method since this diversity encompasses TE types found in other eukaryotes. Furthermore, TEs in *D. melanogaster* are also very often polymorphic and present at low frequency in nature (Charlesworth and Langley 1989) and thus many additional TE insertions exist in natural populations beyond those seen in the reference genome. Finally, the *D. melanogaster* P-element provides a natural control to test our system and compare TSDs and TSMs from natural and artificial insertions since it is absent from the reference genome but widespread in wild populations (Engels, Johnson-Schlitz et al. 1990). Finally, the TSD and TSM from artificial P-element insertions have been extensively characterized (Liao, Rehm et al. 2000; Linheiro and Bergman 2008).

Using resequencing data from 176 isofemale strains of *Drosophila melanogaster* gathered by the *Drosophila* Genetic Reference Panel (DGRP) project (Mackay, Richards et al. 2008), we were able to extract over 11,900 new TE insertions in to over 8,000 new insertion sites. By analyzing data gathered from both Illumina and 454

sequencing platforms, we were able to show that different sequencing platforms give consistent results in terms of TSD discovery. Furthermore, we were able to show that TSDs previously attained through artificial insertions can be comparable to those in natural transposition events. Additionally, we found that TE families from the same subclass present similar TSDs. All target site motifs discovered from DNA and LTR elements were found to be a palindrome whose center varied according to the length of the TSD and whose TSM was an extended version of the TSD. Together these results demonstrate that population genomic resequencing data can be used to rapidly discover TSDs and TSMs in a wild type genomic context allowing a better understanding of TE targeting in nature.

4.3 Materials and Methods

4.3.1 Data origin

Data from the DGRP project was downloaded from NCBI by searching in the Sequence Read Archive (SRA) for DGRP and applying a PERL (v5.8.8) script to summarize the available data (Supplemental file 4.1). Compressed fastq files were downloaded from NCBI 14-16 September 2010. Meta-data for each strain was downloaded summarized (Supplemental file 4.2 and 4.3) and all reads from the same strain were concatenated into fasta files with paired reads from the same fragment with unique identifiers. Data from both the 454 and Illumina platforms were processed separately using identical pipelines.

4.3.2 Identifying *de novo* TE insertions in the DGRP project samples

After sorting the data by sample, we parsed the WGS reads using two selection processes. In the first selection process, we used default settings of BLAT (version 34) (Kent 2002) to query the WGS reads against the TE fasta file that contains canonical sequences for 128 *D. melanogaster* TEs. We kept reads whose best matches included the start (the first 5' end base) or end (the last 3' end nucleotide) of the TE query. Sequences were selected according to the number of blocks matching, number of mismatching and gap bases (allowing 1 mismatch in 25 in both the query and target sequences) and length of the match. When a WGS read had two or more hits for the same TE, we hierarchically selected according to above-mentioned parameters. When a match was indistinguishable between a start/end of TE, we randomly picked one, and when a match was indistinguishable between a start/end and the middle of a TE, we selected for the start/end match. If a WGS read had two or more matches to different TEs, we discarded it if they overlapped and kept the best hit if they did not. During this first selection we got over 6 million sequences that uniquely matched a start or end of a TE. This represents less than 0.1% of the total number of sequences in all the 176 Illumina sequenced strains. For the 454 data, we retained close to 1 million sequences representing 0.5% of the total number of the 454 sequences.

During the second selection process, we mapped reads that included the starts or ends of TEs built from step one to the Release 5 *D. melanogaster* genome sequence using default BLAT settings. We selected for mapped WGS reads with low number of mismatching bases (allowing 1 mismatch in 20 for both the query and target sequences). WGS reads were only selected if a match to the genome or TE included the beginning or end of the read. The selected reads also had to match the reference TE start/end exactly where the genomic region begins or vice versa. Selected WGS reads could only map to the genome in one location. If there was ambiguity about the exact location in the genome of a sequence with the same criterion, the read was discarded.

To find new insertion sites, we selected for sets of mapped reads matching the same reference TE where the distance between the end of one mapped read and the start of the next read found sequentially in the genome was less than or equal to 20 bp. This overlap distance defines the TSD and thus the maximal TSD length that we can discover using this method is 20 bp (Figure 4.1). The TSD was considered *de novo* TE insertion if there was more than two reads on each side of the predicted TSD in the 454 data, and if one or more read supported each side of the TSD for the Illumina data.

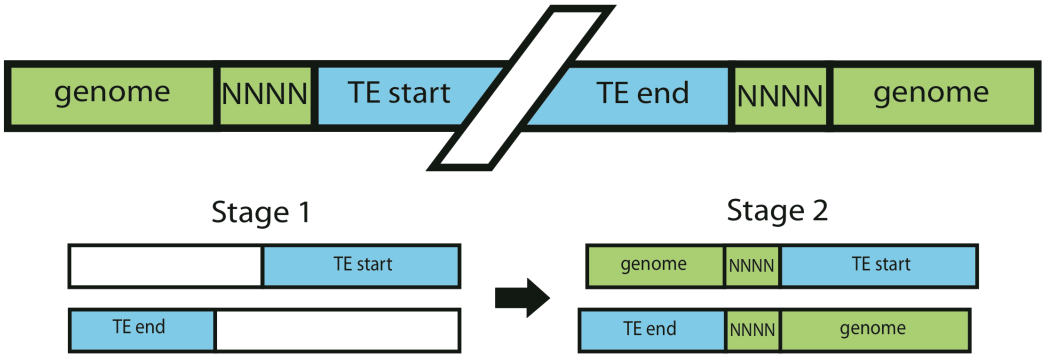


Figure 4.1 Schematic of *de novo* TE insertion site mapping strategy

When a transposable element inserts in to the genome it leaves a characteristic TSD at both its ends as a consequence of its staggered cuts. Top figure is a representation of a full TE mapped to a genome with the TE in light blue and the genomic sequence in green with the TSD indicated by NNNN. Bottom figure represents our two-stage approach to mapping new TEs in the genome. In a first stage we select for reads with a match to either a TE start or TE end and an unmatched part. In the second stage we align the selected reads to the reference genome and look for overlaps, TSDs, between a start and end of a specific TE.

4.3.3 TE logos

Logos were built using an R (version 2.9.1) (R 2009) stand alone implementation of the weblogo algorithm (Crooks, Hon et al. 2004) that permitted to build the different motifs automatically. For each TE family a logo was constructed with the non-redundant sites that mapped to a defined strand. For each insertion site -15 to +15 bp from the start position of the TSD were aligned to produce the logo with insertions in the negative strand reverse complemented before inclusion in the alignment.

4.4 Results

4.4.1 Next generation resequencing data can be used to find *de novo* TE insertions

In order to find *de novo* TEs insertion sites in the *D. melanogaster* genome, we first aligned DRGP Illumina and 454 sequencing reads to the set of known *D. melanogaster* TE reference sequences Table 4.1. Reads that mapped to the start or end of the reference TE were selected and subsequently mapped against the *D. melanogaster* reference genome to find the TE insertion site (see Materials and Methods section above for further details). Since our focus is on discovering new target sites in the genome, and not their allele frequency in the population, we only consider non-redundant insertion sites hereafter.

Table 4.1 Summary of the read data from both platforms

Note that the # of reads mapped to the genome is relative to the number of reads selected in the first and second selection

	Total # of reads	# Mapped TE starts or ends	# Mapped starts	# Mapped ends	# Mapped to the Genome
Illumina	7,835,189,604	6,063,063	3,133,159	2,929,904	97,996
454	209,979,997	956,753	451,423	505,330	48,543

We were able to find insertion sites in all 34 strains sequenced by the 454 platform. However, we were not able to identify insertion sites for 10 out of the 176 Illumina strains analyzed. All of these 10 strains (SRS003467, SRS003469, SRS003470, SRS003474, SRS003475, SRS003476, SRS003486, SRS003487, SRS004126, SRS004137) had read lengths less than 64 bp long. For the 166 strains sequence sequenced by the Illumina platform with data of length greater then 75 bp, we identified a minimum of 20 new insertions per strain with 3 exceptions (SRS003443, SRS003447 and SRS003448) that had fewer than 7 new insertions per strain (see discussion below) (Figure 4.2).

The difference in the numbers of strains sequenced using either platform was reflected in both the number of insertion sites and families that we were able to obtain *de novo* insertion data for. The total number of insertions sites was 8,033 for Illumina and 2,622 for 454. In total, we found *de novo* insertions for 46 families from both platforms, with

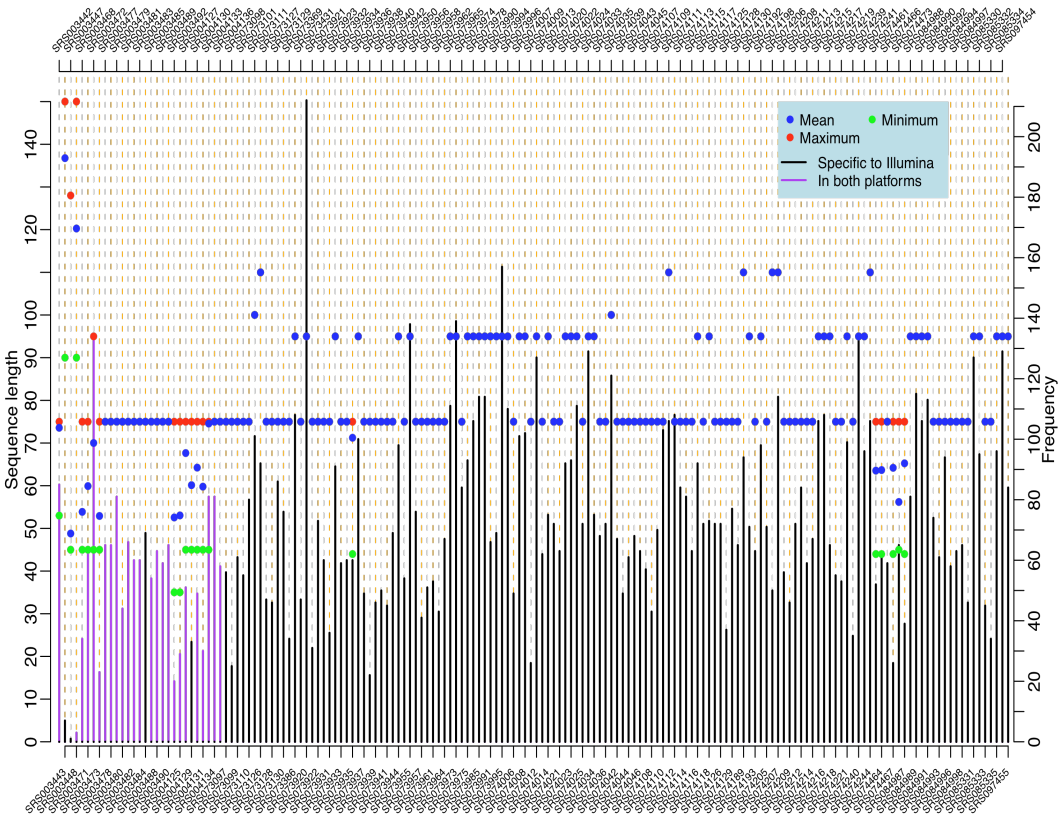


Figure 4.2 Sequence length and number of insertions per strain for the Illumina platform.

Points represent the maximum, minimum and mean of the read length for each of the strains. Bars represent the total number of elements identified per strain. The scale on the left is for the sequence length and the scale on the right is for the number of non-uniquely insertions identified in that strain. The strain labels are alternated on the top and bottom of the graph. Our method attempts to find *de novo* insertions only and thus underestimates the number of TEs per strain.

data from 13 families present only in the Illumina platform and data from 5 families present only in the 454 platform.

We compared the 25 strains that had been sequenced on both platforms and counted the number of times each one of the most abundant TEs (defined as families with more then eight insertion sites in the Illumina dataset) was seen in the same location in the genome for each strain (Table 4.2)**.** In total we found 1,204 insertion sites in Illumina and 1,733 in the 454 platform for these 25 strains, and of those 807 were found in the exact same site (e.g. same location, same strand) on both platforms. When an insertion was in the same site it was also in the same orientation in both platforms with the exception of 1 site out of 123 common sites for the P-element and 1 out of 18 sites for the 1360 element (Table 4.2). On average there were more insertion sites predicted in the common strains for the 454 platform than for Illumina, with 69.32 and 48.16 insertions

per strain respectively. Of the 22 TE families analyzed, only 5 showed a higher percentage in the Illumina platform than in the 454 platform, with 1360 and pogo presenting the highest difference among platforms.

Differences in the set of TE insertions discovered between both platforms can be a consequence of differences in the TSD annotation methods used and differences between read lengths in the 454 and Illumina platforms. When selecting for TSDs in the 454 platform we only considered a true site if there were two or more read matches on both the start and end of the TSD while in the Illumina platform we consider a match with only one read on both sides. Reads from the 454 platform are longer (average of 353.17 bp) than the reads from the Illumina platform (average of 77.6 bp) resulting in both more TSDs identified in the 454 platform and more TSDs identified in the Illumina platform. Longer read length could both promote discovery of insertion sites, with longer reads having more unique matches in the reference genome and TE than smaller reads. Longer reads could however also hinder insertion discovery for old families since divergence between the reference genome and the strain increases the chance of finding divergent sites which can interfer with mapping to the reference genome or TE (see discussion below).

Table 4.2 Comparison of *de novo* TE insertions in 25 strains sequenced by both Illumina and 454 platforms

Note that the number of insertions in the same site also corresponds to the number of insertions with the same TSD.

TE	Non-redundant insertion sites in Illumina	Non-redundant insertion sites in 454	Same site and strand	Same site
hobo	192	282	159	159
1360	52	43	17	18
P-element	148	213	122	123
pogo	160	94	53	53
S-element	1	6	1	1
hopper	79	73	29	29
297	5	9	4	4
412	75	105	53	53
blood	62	87	48	48
Burdock	81	121	63	63
gtwin	3	2	0	0
gypsy	19	20	14	14
HMS-Beagle	58	73	44	44
Mdg1	9	113	6	6

opus	130	219	104	104
Quasimodo	2	2	0	0
Stalker2	15	29	11	11
Tabor	26	44	20	20
Transpac	35	32	21	21
3S18	15	29	11	11
Max-element	18	32	17	17
Roo	19	105	10	10

Despite the fact the 454 data provided more insertions per strain, we chose to focus our analysis on Illumina data since there were more strains, insertions and families from the Illumina platform.

4.4.2 Insertions are spread unevenly through the different classes and subclasses.

Using Illumina resequencing data from 166 strains of *D. melanogaster*, we were able to extract 8,033 non-redundant TE insertions sites, with each strain contributing approximately 48.39% new insertion sites per strain. The DNA transposon class generated the highest number of *de novo* insertions with 4,163 insertion sites spread throughout 5 subclasses and 7 families (Table 4.3 and Table 4.4). The DNA transposon family with the greatest number of new insertion sites is the P-element (n=1,226 insertion sites), a family that is not present in the reference genome sequence (Kaminker, Bergman et al. 2002). The LTR retrotransposon class also generated a large number of *de novo* insertions with a total of 3,861 insertion sites from 3 different subclasses and 31 different families. The LTR subclass with the highest number of insertions was Gypsy with 3,445 insertion sites in 25 different families. The most abundant LTR family was the *opus* element with 1,030 insertion sites. Since our TSD identification strategy requires a fixed TSD length, we were only able to gather a total of 9 new insertion sites for the non-LTR retrotransposon class because of the variable length TSDs from non-LTR elements (Eickbush and Malik 2002). As a consequence of these low numbers, data from non-LTR elements are not considered here further. In total we were able to map *de novo* TE insertions for 41 TE families, 22 of which we found over 8 *de novo* TE insertion sites and have sufficient data to draw conclusions about TSD length and motif properties and are analyzed in further detail here (Table 4.3).

Table 4.3 Number of *de novo* TE insertions per class and subclass identified using Illumina resequencing data.

Class	Insertions per class	Subclasses	Insertions per subclass	Number of families
DNA	4163	hAT	1198	1
		P	1505	2
		Pogo	895	1
		Tc1	25	1
		Transib	540	2
non-LTR	9	I	8	2
		Jockey	1	1
LTR	3861	Copia	1	1
		Gypsy	3445	25
		Pao	415	5
Total	8033		8033	41

4.4.3 Target site duplications have a characteristic length for 30 TE families

We plotted the frequency distribution of TSD lengths for each TE family, and could see a single peak in TSD length for 30 TE families (Figure 4.3; Table 4.4). Although the modal TSD length was often shared by >95% of insertions from a family, there were some cases in which the TSD was different from the majority (Table 4.4). These cases represented a minority of the total number of predicted TSD (1.8%) and were typically only ±1 bp from the optimal TSD for most elements with the exception of opus, which presented a characteristic 2- bp spacing pattern (see discussion below). The modal TSD length for the majority of TE families based on Illumina data agreed both with data from the 454 platform when there was more then 3 insertions per family.

TSDs of the same subclass often, but not always, showed similarities in length, with families showing a different TSD length often having very few insertions. LTR elements from the Gypsy subclass presented a strong preference for a TSD of four bp, with only two out of 25 families (Idefix TSD=5 bp, n=1; Tirant, TSD=2 bp, n=2) showing a different optimal TSD length. For the Pao subclass 3 elements had a TSD of five bp, with the aurora-element (n=2) and rooA (n=1) presenting unusually large TSDs of 16-17 bp and 13 bp, respectively. Both DNA elements from the Transib subclass (hopper and transib2) showed a preference for a TSD of five bp. TEs from the P-

element subclass did not agree with each other concerning the size of TSD, with the P-element presenting a TSD of 8 and 1360 a TSD of 7. There was only one new insertion identified for the Copia subclass from the Dm88 family and this element showed a TSD of 3. On the 454 platform it was possible to gather more insertion sites from this subclass, with the 1731 (n=1) and copia (n=85) families both presenting a TSD of 5 bp.

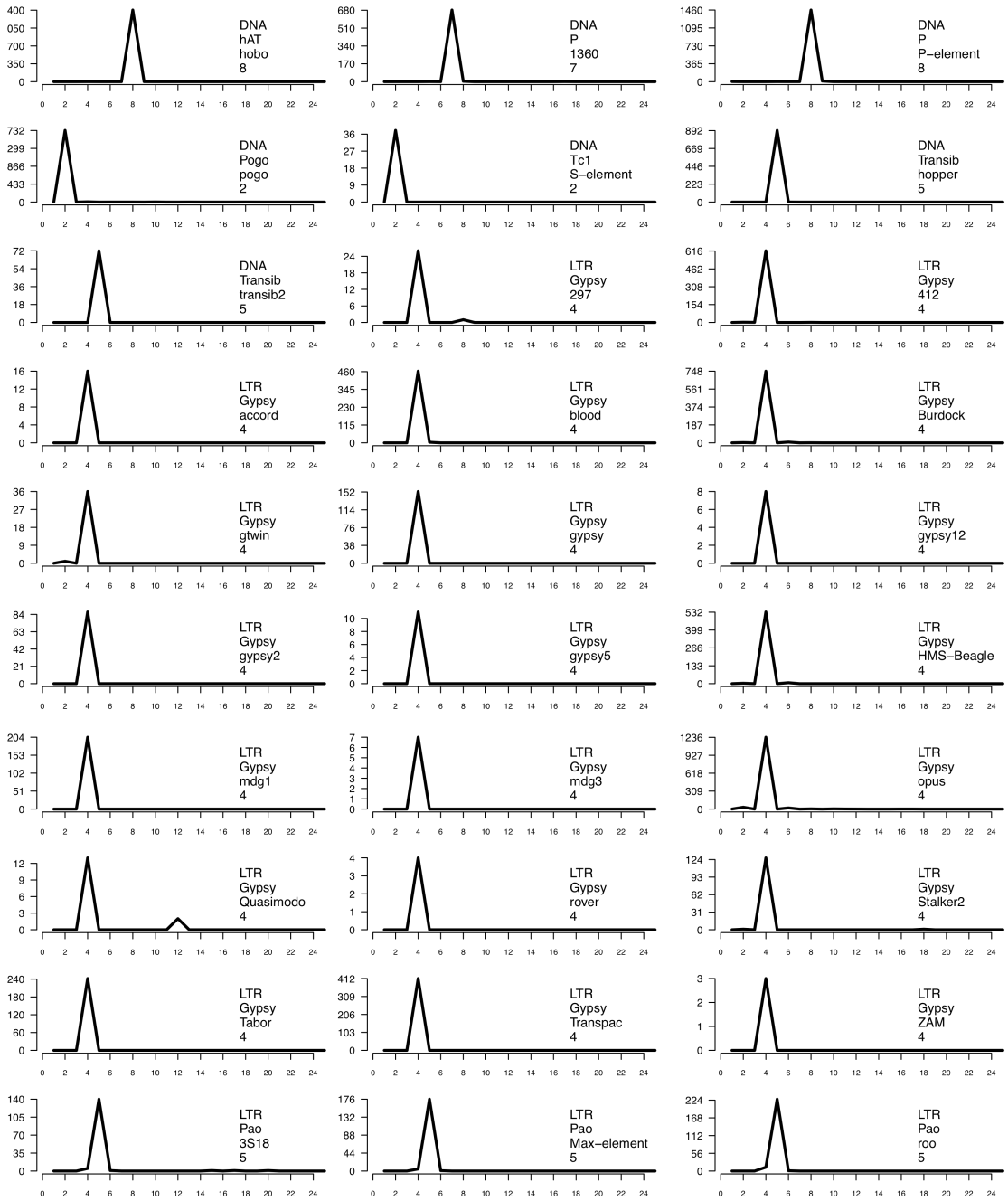


Figure 4.3 Frequency distribution of TSD lengths for different TE families

Predicted TSD lengths for *de novo* insertions are plotted for TE families with 3 or more insertion sites. The plots are organized by class with its class, subclass, name of the TE and the predicted TSD length in the top right corner. All graphs have the same x-axis (from 0 to 25 in windows of 2) with the y-axis varying according to the frequency of the elements. All elements show one major peak with a high level of consistency within the Gypsy and Pao subclasses.

Table 4.4 Number of insertion sites and optimal TSD length based on Illumina data.

Class	Subclass	Family	Insertion sites	Most frequent TSD length	Insertion sites with predicted TSD	% Insertion sites with predicted TSD
DNA	hAT	hobo	1198	8	1196	99.83
DNA	P	1360	279	7	274	98.21
DNA	P	P-element	1226	8	1207	98.45
DNA	Pogo	pogo	895	2	883	98.66
DNA	Tc1	S-element	25	2	25	100
DNA	Transib	hopper	533	5	532	99.81
DNA	Transib	transib2	7	5	7	100
LTR	Copia	Dm88	1	3	1	100
LTR	Gypsy	297	19	4	18	94.74
LTR	Gypsy	412	498	4	494	99.20
LTR	Gypsy	accord	3	4	3	100
LTR	Gypsy	blood	378	4	376	99.47
LTR	Gypsy	Burdock	481	4	471	97.92
LTR	Gypsy	gtwin	19	4	18	94.74
LTR	Gypsy	gypsy	92	4	92	100
LTR	Gypsy	gypsy12	1	4	1	100
LTR	Gypsy	gypsy2	2	4	2	100
LTR	Gypsy	gypsy5	6	4	6	100
LTR	Gypsy	HMS-Beagle	320	4	311	97.19
LTR	Gypsy	Idefix	1	5	1	100
LTR	Gypsy	invader3	1	4	1	100
LTR	Gypsy	invader6	1	4	1	100
LTR	Gypsy	mdg1	146	4	146	100
LTR	Gypsy	mdg3	5	4	5	100
LTR	Gypsy	micropia	1	4	1	100
LTR	Gypsy	opus	1030	4	976	94.76
LTR	Gypsy	Quasimodo	9	4	8	88.89
LTR	Gypsy	rover	3	4	3	100
LTR	Gypsy	Stalker2	84	4	82	97.62
LTR	Gypsy	Tabor	138	4	138	100
LTR	Gypsy	Tirant	2	2	2	100
LTR	Gypsy	Transpac	202	4	202	100
LTR	Gypsy	ZAM	3	4	3	100
LTR	Pao	3S18	119	5	113	94.96
LTR	Pao	aurora-element	2	17-18	2	100
LTR	Pao	Max-element	100	5	96	96.00
LTR	Pao	Roo	193	5	182	94.30
LTR	Pao	rooA	1	13	1	100

4.4.4 Our data corroborates data from previous analysis

To evaluate if our mapping strategy leads to biological realistic TE target site preferences, we compared the TSD data obtained from natural insertions of *D. melanogaster* TEs in the DGRP data to those discovered from artificial P-element insertions in the *Drosophila* Genome Disruption Project (DGDP) and to those previously reported in the literature for 18 out of the 22 families that had more then 8 insertion sites in the DGRP data (Table 4.5). All of the TSD lengths discovered here are consistent with TSD lengths previously observed in the literature.

Table 4.5 TSDs identified in this study compared with previous publications and motifs.

Family	TSD	Previous TSD	Previous TSD Motif	Reference
hobo	8	8	NTNNNNAN	(O'Brochta, Stosic et al. 2009)
1360	7	7	KTNBWAB	(Reiss, Quesneville et al. 2003)
P-element	8	8	GTCCGGAC	(Engels, Johnson-Schlitz et al. 1990; Linheiro and Bergman 2008)
pogo	2	2 or 0	TA	(Tudor, Lobocka et al. 1992)
S-element	2	2	AT	(Merriman, Grimes et al. 1995)
hopper	5	5	N.A.	(Bernstein, Lersch et al. 1995)
transib2	5	5	CABHG	(Kapitonov and Jurka 2003)
297	4	4	N.A.	(Dunsmuir, Brorein et al. 1980)
412	4	4..6	WKRK	(Bowen and McDonald 2001)
blood	4	4	RKAS	(Bowen and McDonald 2001)
Burdock	4	4	TATA	(Bowen and McDonald 2001)
gtwin	4	4	TGTA	(Bowen and McDonald 2001)
gypsy	4	N.A.	N.A.	N.A.
gypsy5	4	4	N.A.	(Bowen and McDonald 2001)
HMS-Beagle	4	4	TRTA	(Bowen and McDonald 2001)
mdg1	4	4	CTAC	(Bowen and McDonald 2001)
opus (nomad)	4	4	TANA	(Whalen and Grigliatti 1998)
Stalker2	4	N.A.	N.A.	N.A.
Tabor	4	4	MMKS	(Bowen and McDonald 2001)
Transpac	4	N.A.	N.A.	N.A.
3S18	5	N.A.	N.A.	N.A.
Max-element	5	N.A.	N.A.	N.A.
roo	5	5	VWWAY	(Bernstein, Lersch et al. 1995)

4.4.5 Target site motifs for DNA and LTR elements are palindromes that share similarity between families in the same TE subclass.

We next identified sequence motifs associated with the TSD by aligning the sequences of insertion sites flanking and including the TSD, which we refer to as target site motifs (TSMs). TSMs can in principle extend beyond the TSD, as has been shown for the P-element (Liao, Rehm et al. 2000; Linheiro and Bergman 2008). A high number of insertions did not necessarily lead to a high information content TSM. Elements with just 25 insertion sites could give a clear motif (S-element) while elements with over 100 insertion sites could originate in a very degenerate motif (Figure 4.4). For 15 TE families, a motif had also been previously reported in the literature that was consistent with our data (Table 4.5). The only motif that had been inferred from more then 10,000 insertion sites was the DNA P-element (Linheiro and Bergman 2008). The only difference between the P-element TSM from natural insertions and the motif from the artificial insertions was an A at position -1 instead of an N, and an N at position -2 replacing an A.

Both DNA and LTR classes of TE showed a preference for a palindromic motif that extended beyond the TSD. In general, we found that the length of the TSM palindrome was dependent on the length of the TSD, and the center of the palindromic TSM varied according to the length of the TSD. When the TSD was an even number, the center of the TSM would between TSD/2 and (TSD/2)+1 bps, counting the insertion nucleotide as the first base and then extending the motif towards the 5' end. When the TSD was an odd number, the center of the palindrome would be at the base located at position (TSD+1)/2 counting the first base as the insertion base. The minimum extension of the TSM for an odd number TSD will be of ±(TSD+1)/2 and for an even TSD will be from ±(TSD-1) from the center of the palindrome. The S-element and Pogo elements did not follow this pattern of TSD extension, with 2*TSD on each side of the palindrome (Figure 4.4).

As with TSD length, TEs from the same subclass showed a similar preference of bases in their TSMs. In the DNA P subclass there was a tendency to have an ANAGT motif that started at position -3, and an ACTNT motif starting at position 5 and 6 for the 1360

and P-element families, respectively. TEs from the Transib subclass (hopper and transib2) insert into a CCANTGG TSM (Figure 4.4). Although the pogo and S-element families were from different subclasses they presented a 2 bp TSD and inserted in to an AT rich region Figure 4.4.

In the LTR class there was also a correspondence between the subclass and the TSM. The Gypsy subclass presented a preference to insert in to a TATATA sequence. The first T at position -1 for the less conserved motifs of 412, blood, mdg1, stalker2 and Tabor and starting at position -2 and extending it with another TA in the more conserved motifs. Exceptions to these rules for the gypsy subclass were the Transpac and 296 families, with a CATATG TSM starting at position -1. The Pao subclass showed a preference to insert in to a weak ATTANNNNANT TSM.

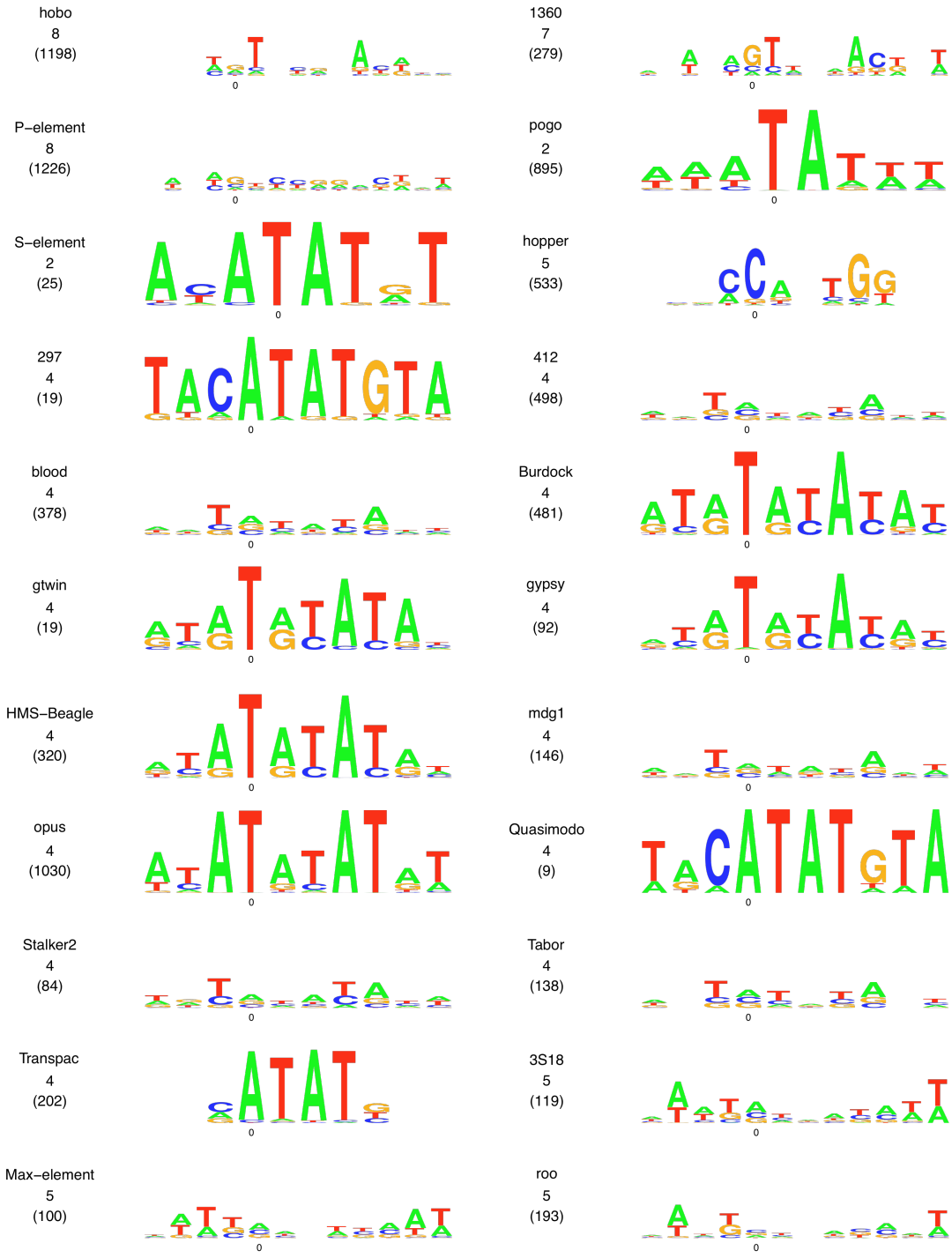


Figure 4.4 TSM logos for TE families with more than eight non-redundant insertion sites.

Numbers in parentheses indicate the number of non-redundant insertion sites used to construct each logo. The y-axis is equal for all the logos and goes up to a bit score of 2. On the x-axis the 0 represents the nucleotide that would be annotated if insertions are mapped to a single nucleotide. All insertions in the negative strand were reverse complemented before constructing the logo. Logos of insertions in just either the positive or negative strand were equal to the ones represented above (Supplemental file 4.4).

4.5 Discussion

In this chapter, we have shown that whole genome shotgun sequence data from next generation sequencing technologies can be used to identify *de novo* TE insertions and discover properties of TE insertion sites. Comparison of TSD lengths and motifs discovered here to those previously reported in the literature demonstrated that this high-throughput strategy leads to results that are consistent with previous small scale studies. Furthermore, TSD data obtained from both Illumina and 454 sequencing platforms also agreed with each other. We also found that the TSD length showed consistency within families and subclasses, all TSMs were a palindrome that was also comparable within families and subclasses.

4.5.1 Limitations of the current approach to finding *de novo* TE insertions

In order to get reliable results, our current approach is very stringent and only allowed for an exact match to the beginning or end of a reference TE at the genome-TE intersection (see methods above). This requirement is the most important factor for the lack of non-LTR elements in our results, since non-LTR elements are often truncated at the 5' end of the TE (Eickbush and Malik 2002) and therefore many *de novo* insertions are filtered out at the very first stages of our analysis. For example, the number of reads with jockey starts is 289 and the number of reads with jockey ends is 6,600. We were however able to gather a reasonable number of insertions for some non-LTR elements like the I-element, which is a recent invader from *D. simulans* (Bucheton, Busseau et al. 2002) and apparently has a higher proportion of non-truncated insertions. One method to overcome the lack of non-LTR 5' ends would be to analyze the observed size of 5' truncations and allow a set number of bases as a tolerance parameter for including non-LTR 5' junction reads. Although this would allow the discovery of more new non-LTR insertions, it would also give rise to many false positives if applied to other TE classes and requires implemention of different methods for different TE classes that would need to be further assessed.

The stringency of our methods also leads to an underestimation in the number of TEs per family in each strain. For example, for the P-element it has been reported that there are about 30 to 50 elements per strain (O'Hare, Driver et al. 1992) while we observe

only around 11 insertions per strain, varying from 1 to 33 in the Illumina platform (Figure 4.5, light blue). Some of this underestimation may come from our methods and some may come from the length of sequencing reads, since the minimum read length from which it was possible to get insertions from was 75 bp long (Figure 4.2). Furthermore, when comparing between 75 and 95 bp long reads on the Illumina platform, it is clear that the number of TEs identified increases with longer reads (Figure 4.5).

Depth of coverage also appears to affect our results. The average coverage of the strains from which we were able to gather insertions was 29 with a mean of 72.1 insertions per strain. The strains with the highest number of insertions were SRS073995 and SRS074240 (134 and 157 insertions each) both with 95 bp long reads and with coverage of 34.5 and 34.4 respectively (Supplemental file 4.2). It is worth mentioning that these two strains also presented 31 P-element insertions each, which is more consistent with previous observations (O'Hare, Driver et al. 1992). To improve our results ideally we would analyze genomes with at least 95 bp long reads with sequence coverage of 35.

An additional factor affecting the number of insertion sites predicted is the sequence quality itself, since there were strains that had very few insertions (below eight) although they had reads of 75 bp or longer. For each of the Illumina sequences selected in the first stage we plotted their quality score for each position. The three strains that had fewer than eight new insertions (SRS003443, SRS003447 and SRS003448) showed a scoring pattern different from the predicted NGS pattern, that is the lowest quality bases at the end of the sequence (Figure 4.6 and Supplemental file 4.5). The pattern of quality scores seemed to indicate an adaptor in the middle of the sequence (Wang, Lin et al. 2009) or that two reads had been concatenated in to one (for example reads of 90 and 150 bp). Such errors could influence the TE/genome junction essential in our analysis.

Finally, TE insertions can be deleterious for the host by causing mutations and to reduce these deleterious effects one strategy is to insert inside another TE (Deininger and Roy-Engel 2002). Since we disregard any TE/genome junction sequence that maps to multiple locations in the genome, this can also lead to an underestimation of the number of insertions per strain. Despite these factors that would tend to underestimate the

number of TEs per strain, they should not directly affect conclusions across different families, since insertion discovery criteria was the same for all TEs.

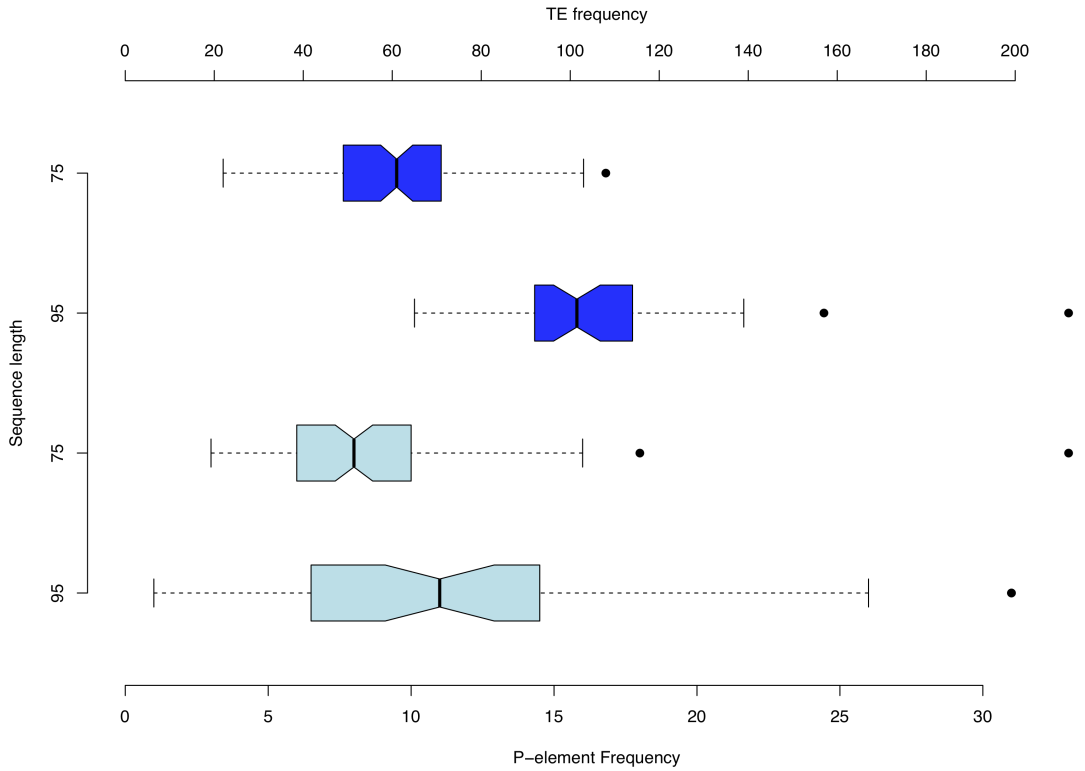


Figure 4.5 Distribution of the number of insertions per strain varies with read length.

The dark blue boxplots are for all insertion sites with their frequency axis above the figure. The light blue boxes are the distribution of the number of P-elements per strain with the axis on the bottom of the figure. The y-axis is divided into strains with 75 and 95 bp long Illumina sequence reads. The notch shows that the mean for the frequency of insertions for the both sequence lengths is significantly larger in the 95 bp long sequences for all TEs and for the P-element.

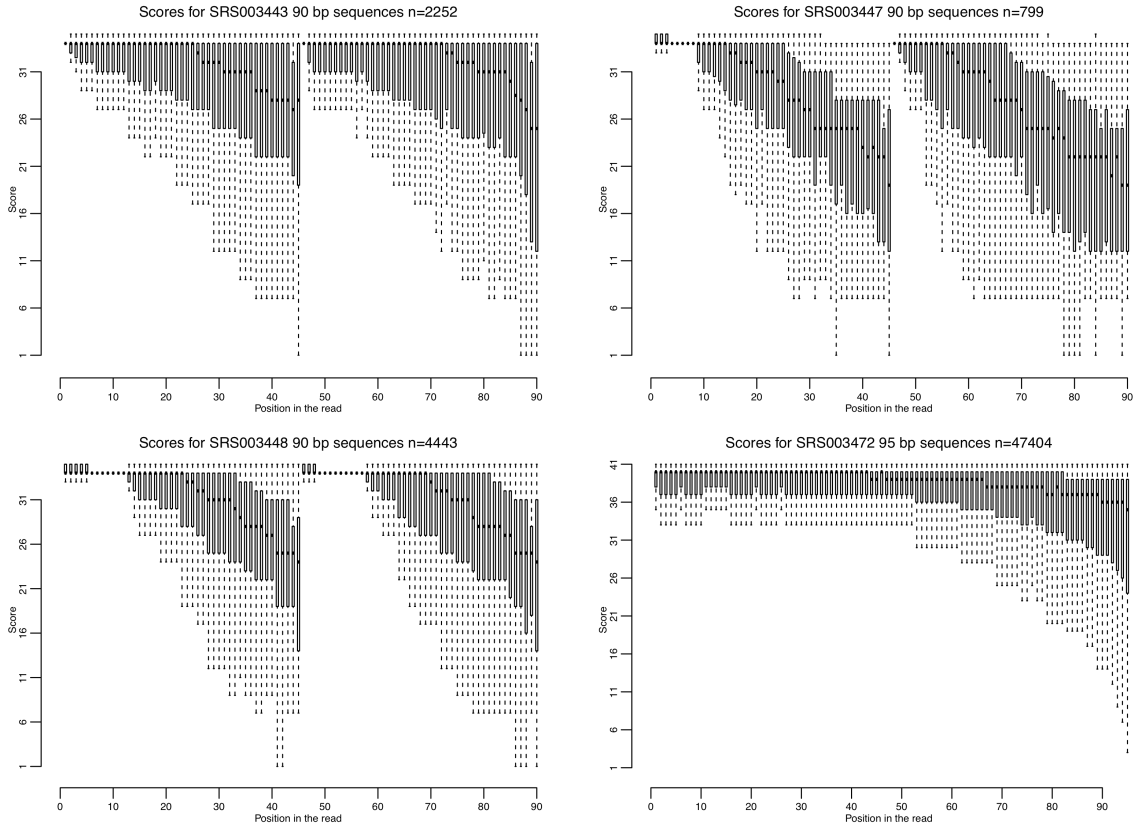


Figure 4.6 Sequence scores according to the position in the read

Shown are the scores boxplots for each position according to strain and read length for the reads that passed our first selection process, the y axis corresponds to the Illumina score (ASCII code -32) and the x axis corresponds to the position in the read with the title indicating the strain, read length and number of reads used. From top to bottom and left to right the first 3 graphs correspond to the 90bp long reads for the 3 strains that had less then 8 insertions each, SRS003443, SRS003447 and SRS003448. The last graph corresponds to a random strain (SRS003472) that presented the expected score pattern for read lengths of 95 bp. The major difference between the strains with a low number of insertions and strain SRS003472 (134 insertions) is a disruption in the scoring pattern in the middle of the read. This pattern looks like two reads of the same length have been put together or that they had 2 adaptors. There are many strains that have 45 and 75 bp reads (Figure 4.2) but only one presented a read length of 64, strain SRS003467.

4.5.2 Other methods for finding TE insertions in next generation sequencing data

Our approach to find *de novo* insertion sites relies on the discovery of the characteristic TSD of each element based on TE/genome junctions in a single sequencing read. This method adds to a number of other approaches that use next generation sequencing to find new TE insertion sites in the genome. One such approach is to have an *a priori* PCR amplification of a specific element and then map the reads to the genome looking for overrepresentations of reads (Ewing and Kazazian 2010; Witherspoon, Xing et al.

2010). These methods have been used to find new LINE elements insertion sites in different genomes (Ewing and Kazazian 2010). These methods require a targeted PCR step, and unless both ends of the TE are amplified, it would not be possible to measure the size of their TSDs since only one side of the TE would be mapped.

Hormozdiari et al (2010) developed a method to use next generation sequencing to find new Alu insertion sites in the human genome that works in a similar way to ours, although the order of alignment processing varied with an alignment first performed to the reference genome and then to the TE. The Hormozdiari et al. (2010) method achieves a high level of coverage for *de novo* insertions through paired end sequence alignments. We do not currently use paired end information, which could potentially give a higher level of coverage. Although by using single reads we are more flexible with our input data, since not all of the strains had paired end reads, and we get solid results without it. Finally, the Hormozdiari et al. (2010) method and the PCR based method mentioned above are optimized and produce results for only one family of TEs, making it less useful for understanding the general processes of transposition in contrast with a more broad-spectrum method such as ours.

There are also methods that instead of looking for *de novo* insertion sites from one or more TE family, focus on the analysis of previously known TE insertion sites in order to compare sequenced strains with reference genomes (Fiston-Lavier, Carrigan et al. 2010). Given an annoation of known TE insertions in a reference genome and for more accuracy the size of TSD, the T-lex method used pair-end Illumina sequences to map the presence or absence of known TE insertions in a resequenced strain. In our approach, we did not find known TEs present in the reference genome, since these do not provide new information about TE insertion site preferences. In principle, we could apply the same technique as we used for detecting *de novo* TEs if we first masked the reference genome for repeats.

4.5.3 Age of TEs may affect *de novo* TE insertion discovery

Although we did not specifically attempt to estimate the allele frequency of each insertion across the entire set of strains in this study, we can get some estimate of whether an individual insertion is rare or common based on whether an insertion site is

found in multiple strains. Interestingly, when we compared 454 and Illumina platforms, we could see differences across platforms for TE discovery in terms of age of TE. The families with the highest number of non-redundant (i.e. young) insertion sites (hobo, P-element, blood, roo and 412) all had at least 40% more insertions in the 454 platform than in Illumina **(**Table 4.2). The families that had more redundant (i.e. old) insertion sites (1360, pogo, Transpac and gtwin) showed the opposite pattern, with at least 8.5% fewer insertions in 454 than in Illumina (Table 4.2).

For families with a high number of young insertion, we could see a higher number of insertions in the 454 platform because of the conservation of the TE end sequences, which would lead to the increased detection rates in longer 454 reads. For families with older elements that are more distant to the reference sequence, it would be easier to identify the conserved ends of the elements when the size of the sequencing read is smaller like on the Illumina platform, since there will be less chance that we will find a mutation between the genomic copy and the reference sequence. To overcome a decrease in insertions on the 454 platform for the more divergent TEs one method would be to allow a family-specific threshold in the number of mismatches to the reference TE.

4.5.4 Both DNA and LTR elements share a preference for palindromic target sequences.

When attempting to extract a target site sequence after an insertion occurred, the TSD can be assumed as the totality of the target site since there is no previous knowledge of the sequence prior to insertion. In our study we were able to look at the genome before and after the TE insertion, allowing us to extend the TE target site on both the 5' end of the insertion and on the 3' end of the insertion (Figure 4.3, Table 4.5 and Supplemental file 4.4). Using our approach, we were able to generate TSMs for a larger number of TE families that established a preference for a palindromic sequence for both DNA and LTR classes, with the structure of the palindrome only varying according to the TSD length.

The preference for a palindromic target site and a consistent structure between both the DNA and LTR classes is an interesting feature that shows an overlap between the

transposase and retrotransposase activities. It is known that some transposase and retrotransposase proteins work in dimeric or multimeric units that may or not be equal (Eickbush and Malik 2002; Tang, Cecconi et al. 2007). Given the similarity between the target structures both types of transposases might work similarly, with different units attacking the same half target site on opposite strands. The fact that when the TSD is an odd number of bases the palindrome has an undefined base at its center can be connected to the structure of a palindrome.. When reading the same sequence from the positive and negative strands it would be impossible to get the exact same sequence on both sides since a base can not be the complement of itself, cancelling any overrepresentation of a specific base. Therefore if a TSD is an odd number of bases the middle base of the palindrome at the target site would be predicted not have a defined base, as observed.

4.5.5 Some LTR and DNA families show multiple TSD lengths.

Although most families showed a consistent TSD length, the LTR element *opus* from the gypsy subclass had nearly 5% of its insertion sites (54) presenting a TSD different from the predicted 4 bp. One interesting aspect in the different TSDs of this element was that non-optimal predicted TSD had lengths of 2, 6, 8 and 10 bp respectively with none presenting a TSD of 3, 5, 7 or 9 bp. Although with fewer insertion sites with a different TSD from the optimum, the DNA element *pogo* also demonstrated a slight preference for even number TSDs with 2, 3, 1, 1, 1 and 3 of its insertions sites being 1, 4, 6, 7, 8 and 10 bp long respectively. In contrast to the elements with a short TSD of 2 and 4 bp, the elements with a TSD of 5 bp and above presented most of its non-optimal TSDs close to the optimal one (±1 bp). The preference for a TSD of an even number in some elements with a TSD of 2 and 4 might be related with the fact that these sites are AT rich sequences with the transposase occasionally mis-cutting at the end of a nearby A or T, showing an important mechanistic interaction between the transposase and the target sequence.

5 Conclusions and Future Work

This section summarizes the major findings in this thesis that were covered in the previous chapters and discusses the contribution of results presented in this thesis to the knowledge of P-element targeting. Some possible future work to understand what is still missing about the P-element insertion mechanism will also be discussed in this final chapter alongside with some preliminary results attained from the comparison of data from chapter 3 and 4 to address these open questions.

5.1 Conclusions

The P-element is a unique TE system for understanding the understanding TE targeting due to its recent invasion in the *Drosophila melanogaster* genome (Daniels, Peterson et al. 1990), which has led to the generation of large a data set of P-element insertions and a a large body of knowledge about its transposition mechanisms. In order to understand its targeting mechanism in better detail, in Chapter 2 we studied the patterns of some of the most simple aspects of P-element insertion such as distance between insertions and frequency of insertion in each strand. We found patterns in the distance between insertions and the random usage of each strand that could be understood by the structure of the P-element 14 bp palindromic target site motif. Additionally we found that the three bp that flanked the core of the target on either side of the TSD were destroyed by P-element insertion but then complemented by the P-element ends. Although we could explain several aspects of the P-element insertion mechanism and also clarify the genome mappings of distinct artificial P-element families, the information we learned about local P-element target preference could not explain its preference to insert into promoter regions. The inferred TSM was very degenerate and was spread through out the *D. melanogaster* genome. We concluded that it was not the P-element TSM that was causing the pattern of promoter targeting that was the main aim of this thesis, so we directed our analysis to the genomic features located at the 5' end of the genes that might influence P-element targeting.

Thanks to the recent availability of a large number of genome-wide datasets on gene structure and protein binding, in Chapter 3 we were able to investigate some of the structural aspects of promoters that could in turn be linked with P-element targeting. For

an insertion to occur, the transposase needs access to DNA, and one major aspect determining accessibility is the location of the nucleosomes. We therefore studied the correlation between P-element insertions and nucleosome positioning and found that the optimal zone for P-element insertions, from -190 to +80 bp from the TSS corresponded to the nucleosome free region close to the TSS (Mavrich, Jiang et al. 2008), and that there was a negative correlation between insertions and nucleosome density. The discovery that P-elements prefer to insert into nucleosome free DNA explained the positioning of P-element insertions close to the TSS, changes in location of P-element insertion with polymerase activity and association of P-element insertions with genes that have paused RNA polymerase. However, since nucleosome avoidance was also found for other TEs like *piggyBac* and *Minos*, and may be a general feature of transposable elements (Gangadharan, Mularoni et al. 2010), we found that this explanation did not clarify the full determinants of P-element promoter targeting preferences.

We therefore explored other characteristics of promoter architecture that could explain P-element targeting. Once again, we looked for an explanation at the nucleotide level in promoter sequences since the P-element target identified in Chapter 2 was a CG rich palindromic motif [ATRGTCCGGACWAT]. We built a set of predicted promoter motifs for all TSS in *D. melanogaster* and cross-referenced them with P-element insertions. We found, as predicted, a negative association between P-element insertion with promoters that contain the AT rich motif TATA box and a positive association those that have the CG rich motif Downstream Promoter Element (DPE) and the palindromic motif DNA Replication Element (DRE). This investigation did not reveal a strict preference for certain promoter types, since some TSS had neither a DPE or DRE motif but yet were still targeted, and others with at least one of these motifs were not targeted. We then investigated the role of other general transcription associated factors such as the TATA box binding protein (TBP) and the TBP-related factor 2 (TRF2) to widen our understanding of P-element targeting. Although we found a positive association between the P-element and the binding of both of these general transcription factors P-element targeting specificity was not fully explained by either. There were insertions into promoters with both TBP and TRF2 and there were many insertions in to promoters without either of them. We then explored data from a recent publication that investigated the binding patterns of Polycomb Group (PcG) and Trithorax Group (trxG)

proteins across the genome to investiage if these factors influence P-element targeting (Schuettengruber, Ganapathi et al. 2009). By cross-referencing the PcG and trxG data with the P-element targeted TSSs, we were able to arrive at several new insights on P-element targeting since the presence of PcG recruiters and trxG binding could explain 90% of P-element insertions close to the TSSs. We could also relate this strong association between PcG and trxG binding and P-element promoter targetting with the phenomenon of P-element homing since known gene targets of homing are associated with PcG binding (Kassis, VanSickle et al. 1991; Kassis, Noll et al. 1992; Taillebourg and Dura 1999; Bender and Hudson 2000). The strong association with the recruiter PcG and trxG proteins alongside with the nucleosome displacing close to TSS was not sufficient to explain targeting since not all PcG/trxG associated sites were targeted by a P-element and vice versa. Thus, we conclude that P-element promoter targeting is caused by multiple factors including PcG/trxG binding and nucleosome avoidance.

In Chapters 2 and 3, we only looked at the target site preferences of artificial P-element insertions. To understand the target preferences of naturally occurring P-element insertions and other TEs families, in Chapter 4 we used available genome resequencing data from the *Drosophila* Genetic Reference Panel (DGRP) (Mackay, Richards et al. 2008) to find *de novo* TE insertion sites in wild strains of *D. melanogaster*. We found new insertion sites for different families of naturally occurring TEs by searching for hallmarks of their target site duplications (TSD) in the sequence data. We analyzed several different aspects of target site preferences such as the size of the TSD and the target site motif (TSM) for all TE families with sufficient data. We found that the overall number of insertions discovered was an underestimation due to the stringency of our selection criteria and differences in the sequencing platform used. One of the most interesting results of this analysis was the consistency within families concerning the size of the TSD and structure of the TSM for elements from the same subclass. Another discovery was that a general feature of TE target site sequences was the preference for palindromic sites. The preference for palindromic targets was observed for both DNA and LTR elements. Due to the palindromic structure, we found that even-length TSD presented a continuous palindrome, while transposable elements with an odd TSD had a palindrome that was disrupted in the middle. This analysis showed that a preference for palindromic sites is a more general feature of TEs, but that there is a wide diversity of sequence and size for target site motifs among different TE families.

5.2 Future work

One of the major limitations of the work in Chapters 2 and 3 is that we have only analyzed P-element insertions from laboratory experiments whose main objective was the disruption and analysis of gene expression patterns. These data was therefore skewed towards a non-redundant set of genes and potentially presents a biased picture of the true P-element insertion pattern. Therefore, one of the key questions to ask in any further analysis of P-element targeting is if the results presented in this thesis are observed for P-element insertions that arose in a natural environment or if they are unique to laboratory manipulated stocks. Although the full answer to this question is beyond the scope of this thesis, we present here some preliminary results on this issue.

5.2.1 The P element prefers the 5' end of genes with PcG recruiter binding in natural strains.

Using data on *de novo* insertions from all the different natural families that in Chapter 4, we found that for families that had more then eight insertion sites, only the P-element presented a tendency to insert near the TSS (Figure 5.1). We found that over 67% of all natural insertions are located in a ±1Kb window around the TSS as observed for artificial P-element insertions implying a natural preference for TSSs (Figure 3.4A). In Chapter 3, we found that recruiter PcG and trxG binding in promoter regions was positively associated with the P-element insertion. We observe that natural P-element insertions also prefer PcG bound promoters (Figure 5.1, Table 5.1). This analysis of natural P-element insertions, also revealed a preference to insert close to the TSS in the absence of recruiter PcG (Figure 5.1), which should be investigated further in the artificial P-element insertion data.

To see if other natural transposable elements (TEs) could also present this pattern we compared the overlap of all predicted TSDs from *de novo* insertions with both the PcG and trithorax group (trxG) of proteins. To do this we used similar methods to those used in Chapter 3 to get the overlap between each protein group and the TEs and also the overall genome coverage. When analyzing the association between recruiter PcG and TEs the family that showed the strongest positive association (P-value for the binomial test below 2.16E-16 for a 5% expected overlap) was the P-element with over 69%

(851/1,225) of its insertions overlapping recruiter PcG sites. From the remaining TEs the most interesting association was with the 1360 element, another family from the P-element subclass, which also showed a strong positive association with a 15% expected overlap (Table 5.1)

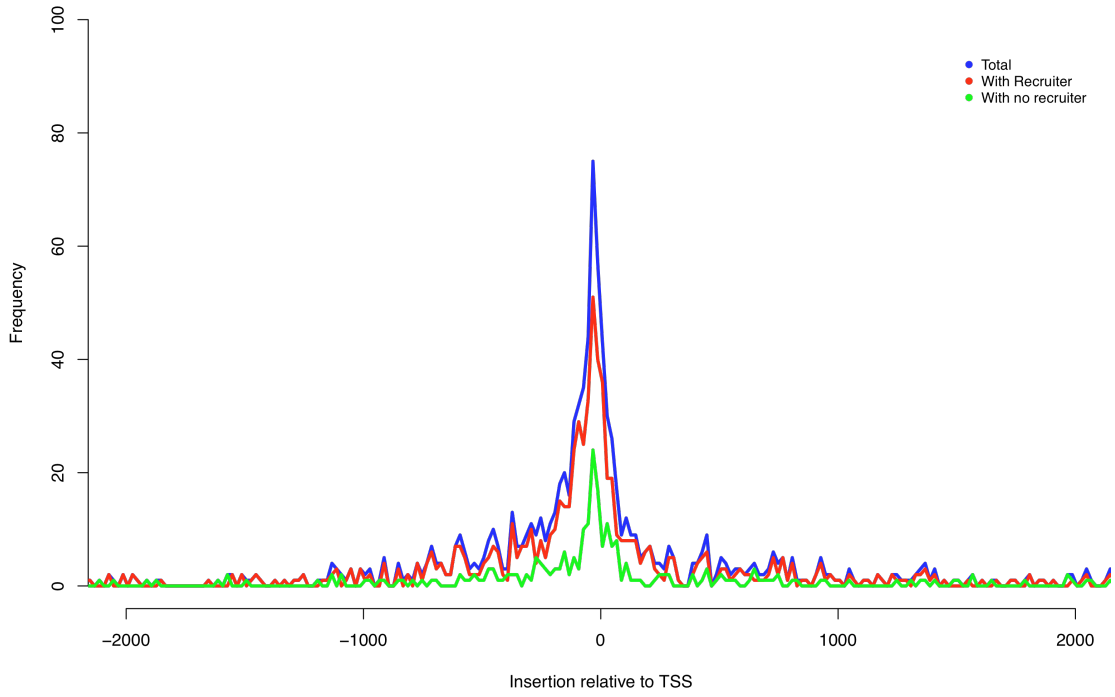


Figure 5.1 Distribution of natural P-element insertions surrounding the TSS.

Frequency of P-element insertions in a ±2 Kb window around the TSS in 20 bp breaks. Blue represents the total number of P-elements surrounding the TSS; red represents P-element insertions in promoters bound by recruiter PcG factors; green represents P-element insertions in promoters that do not overlap a recruiter PcG binding region. This pattern of insertions, skewed towards the 5' end of genes in a ±1Kb window, is very similar to the one for artificial P-element insertions presented in Figure 3.4A. Although there are a higher number of insertion sites close to the TSS for the P-elements that overlap the recruiter PcG the pattern of insertion seems to be conserved in the absence of recruiter PcG binding.

Table 5.1 Association between natural TE insertion from multiple families and recruiter PcG

Element	# with recruiter PcG	% with recruiter PcG	P-Value	Association
hobo	51	4.26	1.53E-01	-
1360	42	15.05	8.65E-10	+
P-element	851	69.41	<2.16E-16	+
pogo	61	6.82	3.53E-02	+
S-element	5	20.00	8.57E-03	+
hopper	41	7.69	1.45E-02	+
297	2	10.53	2.61E-01	+
412	10	2.01	5.42E-04	-

blood	10	2.65	2.03E-02	-
Burdock	46	9.56	1.23E-04	+
gtwin	2	10.53	2.61E-01	+
gypsy	14	15.22	2.98E-04	+
HMS-Beagle	17	5.31	9.00E-01	+
mdg1	2	1.37	3.74E-02	-
opus	22	2.14	8.05E-07	-
Quasimodo	2	22.22	7.68E-02	+
Stalker2	1	1.19	1.34E-01	-
Tabor	5	3.62	5.63E-01	-
Transpac	10	4.95	1.00E+00	-
3S18	5	4.20	8.36E-01	-
Max-element	5	5.00	1.00E+00	-
roo	2	1.04	5.05E-03	-

We repeated the same analysis for trxG bound regions and found that the P-element was the only family with a strong positive association with trxG binding (P-value for the binomial test below 2.16E-16 for a 9% genome coverage). Although this analysis revealed a strong association between trxG bound regions and natural P-element insertion, follow-up analysis suggested that it might be dependent on the overlap between the recruiter PcG and trxG. Both groups overlap each other substantially, with 31.5% (3,500,674 bp/11,109,571 bp) of the trxG sites overlapping the recruiter PcG and 55.7% (3,500,674 bp/6,285,015 bp) of the recruiter PcG sites overlapping trxG. More than 38.1% of the unique P-element insertions sites overlapped a region bound by trxG, and of those 83.1% also overlapped a protein from the recruiter PcG group (P< 2.16E-16 for binomial test with 31.5% overlap). Of the 69% P-element insertions sites overlapping a recruiter PcG, 45.6% also overlapped a trxG protein (P=3.97E-09 for binomial test with 55.7% overlap). Thus the P-element and the recruiter PcG proteins therefore seem to share more of the same sites, while the trxG proteins seem to be associated with the P-element due to their association with the PcG recruiters, which are thought to recruit them to their binding sites (Schuettengruber, Ganapathi et al. 2009). The possibility that there is a direct positive interaction between the P-element and the recruiter PcG proteins and an indirect interaction with the trxG proteins revealed from the natural insertions suggest that this effect should be investigated further in the artificial insertion data as well.

5.2.2 Are P-element insertions associated with GAF binding?

Understanding which of the recruiter PcG factors is most important for P-element promoter targeting is also an open question. If we analyze the different proteins from the recruiter PcG individually and count the number of natural P-element sites that they overlap we can see a strong increase in the number of sites that overlap GAF (Table 5.2). Furthermore, of the 851 unique P-element sites that were associated with a recruiter protein, 706 into regions bound by GAF, and of those 289 were only bound by GAF. The association between the P-element and the GAF recruiter protein could explain the preference for nucleosome avoidance near the TSS, since GAF may recruit the PcG by nucleosome disruption (Schuettengruber, Ganapathi et al. 2009). The role of GAF should therefore be investigated further in studies on P-element promoter targeting.

Table 5.2 Relative coverage of P-element insertions in the recruiter PcG proteins

Note: genome coverage percentage for each protein in the recruiter PcG is indicated in parenthesis

	DSP1 (1.365%)	GAF (2.860%)	PHO (2.248%)	PHOL (1.987%)
Frequency	360	706	264	355
Percentage	21.365	41.899	15.668	21.068
P Value	<2.16e-16	<2.16e-16	<2.16e-16	<2.16e-16

5.2.3 Unanswered questions about P-element target preferences.

Although this thesis provides an improvement in our knowledge of P-element target site selection, there are still many questions that remain answered. Why are P-element insertions associated with recruiter PcG binding? Is GAF the major factor driving association of P-elements with promoters? What happens in the insertion sites where there is no PcG protein? To try and clarify these questions additional research is needed. One way to approach these issues would be to do an all genome-wide ChIP analysis of GAF during the full period of development (from 8 to 24 hours) that the P-element is active (Engels and Preston 1979). This might allow for the discovery of interactions between the P-element and the GAF protein not seen in the 4-12 hours embryos from (Schuettengruber, Ganapathi et al. 2009). It could also be interesting to look at the association of P-element insertion with other proteins such has Pipsqueak (PSQ), Zeste,

Grainyhead (GH) and Sp1/KLF that are also associated in the recruitment of PcG (Schuettengruber, Ganapathi et al. 2009). To understand the nature of the interaction between the GAF protein and the P-element transposase, a thorough examination of both proteins is needed including analysis of their domain structures and potential protein-protein interactions. If the nature of the interaction between both proteins could be deciphered it would open new pathways for the "reprogramming" of the P-element in order to induce insertion into genes that are not currently targeted and continue to allow the P-element to be one of the main tools of *Drosophila* genetics.

6 Bibliography

Aerts, S., G. Thijs, et al. (2004). "Comprehensive analysis of the base composition around the transcription start site in Metazoa." BMC Genomics **5**(1): 34.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Ashburner, M., K. G. Golic, et al. (2005). Transgenesis and the manipulation of genes and gene expression. *Drosophila*: a laboratory handbook. C. S. H. L. Press. New York**:** 311-400.

Bainton, R., P. Gamas, et al. (1991). "*Tn7* transposition in vitro proceeds through an excised transposon intermediate generated by staggered breaks in DNA." Cell **65**(5): 805-16.

Bajic, V. B., S. L. Tan, et al. (2006). "Mice and men: their promoter properties." PLoS Genet **2**(4): e54.

Beall, E. L. and D. C. Rio (1997). "*Drosophila* P-element transposase is a novel site-specific endonuclease." Genes Dev **11**(16): 2137-51.

Beall, E. L. and D. C. Rio (1998). "Transposase makes critical contacts with, and is stimulated by, single-stranded DNA at the P element termini in vitro." Embo J **17**(7): 2122-36.

Bellen, H. J. (1999). "Ten years of enhancer detection: lessons from the fly." Plant Cell **11**(12): 2271-81.

Bellen, H. J., S. Kooyer, et al. (1992). "The *Drosophila* couch potato protein is expressed in nuclei of peripheral neuronal precursors and shows homology to RNA-binding proteins." Genes Dev **6**(11): 2125-36.

Bellen, H. J., R. W. Levis, et al. (2004). "The BDGP gene disruption project: single transposon insertions associated with 40% of *Drosophila* genes." Genetics **167**(2): 761-81.

Bellen, H. J., C. J. O'Kane, et al. (1989). "P-element-mediated enhancer detection: a versatile method to study development in *Drosophila*." Genes Dev **3**(9): 1288-300.

Bender, J. and N. Kleckner (1992). "Tn*10* insertion specificity is strongly dependent upon sequences immediately adjacent to the target-site consensus sequence." Proc Natl Acad Sci U S A **89**(17): 7996-8000.

Bender, W. and A. Hudson (2000). "P element homing to the *Drosophila* bithorax complex." Development **127**(18): 3981-92.

Berg, C. A. and A. C. Spradling (1991). "Studies on the rate and site-specificity of P element transposition." Genetics **127**(3): 515-24.

Bernstein, M., R. A. Lersch, et al. (1995). "Transposon insertions causing constitutive Sex-lethal activity in *Drosophila* melanogaster affect Sxl sex-specific transcript splicing." Genetics **139**(2): 631-48.

Biemont, C. and C. Vieira (2006). "Genetics: junk DNA as an evolutionary force." Nature **443**(7111): 521-4.

Bowen, N. J. and J. F. McDonald (2001). "*Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside." Genome Res **11**(9): 1527-40.

Bownes, M. (1990). "Preferential insertion of P elements into genes expressed in the germ-line of *Drosophila melanogaster*." Mol Gen Genet **222**(2-3): 457-60.

Brand, A. H. and N. Perrimon (1993). "Targeted gene expression as a means of altering cell fates and generating dominant phenotypes." Development **118**: 401-415.
Bucheton, A., I. Busseau, et al. (2002). Mobile DNA II. Washington, D.C., ASM Press**:** xviii, 1204 p., [32] p. of plates.
Carlson, C. M., A. J. Dupuy, et al. (2003). "Transposon mutagenesis of the mouse germline." Genetics **165**(1): 243-56.
Celniker, S. E., D. A. Wheeler, et al. (2002). "Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence." Genome Biol **3**(12): RESEARCH0079.
Charlesworth, B. and C. H. Langley (1989). "The population genetics of *Drosophila* transposable elements." Annu Rev Genet **23**: 251-87.
Chintapalli, V. R., J. Wang, et al. (2007). "Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease." Nat Genet **39**(6): 715-20.
Clark, J. B. and M. G. Kidwell (1997). "A phylogenetic perspective on P transposable element evolution in *Drosophila*." Proc Natl Acad Sci U S A **94**(21): 11428-33.
Collins, J. J. and P. Anderson (1994). "The Tc5 family of transposable elements in Caenorhabditis elegans." Genetics **137**(3): 771-81.
Craig, N. L. (2002). Mobile DNA: an Introduction. Mobile DNA II. N. L. Craig. Washington, D.C., ASM Press**:** 3-11.
Crooks, G. E., G. Hon, et al. (2004). "WebLogo: a sequence logo generator." Genome Res **14**(6): 1188-90.
Cutter, A. D., J. M. Good, et al. (2005). "Transposable element orientation bias in the *Drosophila melanogaster* genome." J Mol Evol **61**(6): 733-41.
Daniels, S. B., K. R. Peterson, et al. (1990). "Evidence for horizontal transmission of the P transposable element between *Drosophila* species." Genetics **124**(2): 339-55.
Davies, C. J. and C. A. Hutchison, 3rd (1995). "Insertion site specificity of the transposon Tn3." Nucleic Acids Res **23**(3): 507-14.
Deininger, P. and A. Roy-Engel (2002). Mobile Elements in Animals and Plants. Mobile DNA II. N. Craig. Washington, D.C., ASM Press**:** 1074-1092.
Dietrich, C. R., F. Cui, et al. (2002). "Maize *Mu* transposons are targeted to the 5' untranslated region of the *gl8* gene and sequences flanking *Mu* target-site duplications exhibit nonrandom nucleotide composition throughout the genome." Genetics **160**(2): 697-716.
Down, T. A., C. M. Bergman, et al. (2007). "Large-Scale Discovery of Promoter Motifs in *Drosophila melanogaster*." PLoS Comput Biol **3**(1): e7.
Drysdale, R. A. and M. A. Crosby (2005). "FlyBase: genes and gene models." Nucleic Acids Res **33**(Database issue): D390-5.
Dunsmuir, P., W. J. Brorein, Jr., et al. (1980). "Insertion of the *Drosophila* transposable element copia generates a 5 base pair duplication." Cell **21**(2): 575-9.
Eggleston, W. B. (1990). P element transposition and excision in *Drosophila*: Interactions between elements. Genetics. Wisconsin, University of Wisconsin. **PhD**.
Eickbush, T. H. and H. S. Malik (2002). Origins and Evolution of Retrotransposons. Mobile DNA II. N. Craig. Washington, D.C., ASM Press**:** 1111-1144.
Ellinghaus, D., S. Kurtz, et al. (2008). "LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons." BMC Bioinformatics **9**: 18.
Engels, W. R., D. M. Johnson-Schlitz, et al. (1990). "High-frequency P element loss in *Drosophila* is homolog dependent." Cell **62**(3): 515-25.

Engels, W. R. and C. R. Preston (1979). "Hybrid dysgenesis in *Drosophila melanogaster*: the biology of female and male sterility." Genetics **92**(1): 161-74.
Ewing, A. D. and H. H. Kazazian, Jr. (2010). "High-throughput sequencing reveals extensive variation in human-specific *L1* content in individual human genomes." Genome Res **20**(9): 1262-70.
Fawcett, D. H., C. K. Lister, et al. (1986). "Transposable elements controlling I-R hybrid dysgenesis in *D. melanogaster* are similar to mammalian LINEs." Cell **47**(6): 1007-15.
Fernandes, J., Q. Dong, et al. (2004). "Genome-wide mutagenesis of *Zea mays* L. using *RescueMu* transposons." Genome Biol **5**(10): R82.
Fiston-Lavier, A. S., M. Carrigan, et al. (2010). "T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data." Nucleic Acids Res.
FitzGerald, P. C., D. Sturgill, et al. (2006). "Comparative genomics of *Drosophila* and human core promoters." Genome Biol **7**(7): R53.
Fontanillas, P., D. L. Hartl, et al. (2007). "Genome organization and gene expression shape the transposable element distribution in the *Drosophila melanogaster* euchromatin." PLoS Genet **3**(11): e210.
Fuda, N. J., M. B. Ardehali, et al. (2009). "Defining mechanisms that regulate RNA polymerase II transcription in vivo." Nature **461**(7261): 186-92.
Gangadharan, S., L. Mularoni, et al. (2010). "Inaugural Article: DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo." Proc Natl Acad Sci U S A **107**(51): 21966-72.
Garrell, J. and J. Modolell (1990). "The *Drosophila* extramacrochaetae locus, an antagonist of proneural genes that, like these genes, encodes a helix-loop-helix protein." Cell **61**(1): 39-48.
Gershenzon, N. I., E. N. Trifonov, et al. (2006). "The features of *Drosophila* core promoters revealed by statistical analysis." BMC Genomics **7**: 161.
Geurts, A. M., C. S. Hackett, et al. (2006). "Structure-based prediction of insertion-site preferences of transposons into chromosomes." Nucleic Acids Res **34**(9): 2803-11.
Gilmour, D. S. (2009). "Promoter proximal pausing on genes in metazoans." Chromosoma **118**(1): 1-10.
Gloor, G. B., N. A. Nassif, et al. (1991). "Targeted gene replacement in *Drosophila* via P element-induced gap repair." Science **253**(5024): 1110-7.
Golic, K. G. and M. M. Golic (1996). "Engineering the *Drosophila* genome: chromosome rearrangements by design." Genetics **144**(4): 1693-711.
Golic, K. G. and S. Lindquist (1989). "The FLP recombinase of yeast catalyzes site-specific recombination in the *Drosophila* genome." Cell **59**(3): 499-509.
Goryshin, I. Y., J. A. Miller, et al. (1998). "Tn*5*/IS*50* target recognition." Proc Natl Acad Sci U S A **95**(18): 10716-21.
Granger, L., E. Martin, et al. (2004). "*Mos* as a tool for genome-wide insertional mutagenesis in *Caenorhabditis elegans*: results of a pilot study." Nucleic Acids Res **32**(14): e117.
Green, M. M. (1977). "Genetic instability in *Drosophila melanogaster*: De novo induction of putative insertion mutations." Proc Natl Acad Sci U S A **74**(8): 3490-3493.
Guimond, N., D. K. Bideshi, et al. (2003). "Patterns of *Hermes* transposition in *Drosophila melanogaster*." Mol Genet Genomics **268**(6): 779-90.

Hallet, B., R. Rezsohazy, et al. (1994). "IS*231A* insertion specificity: consensus sequence and DNA bending at the target site." Mol Microbiol **14**(1): 131-9.

Halling, S. M. and N. Kleckner (1982). "A symmetrical six-base-pair target site sequence determines Tn*10* insertion specificity." Cell **28**(1): 155-63.

Hammer, S. E., S. Strehl, et al. (2005). "Homologs of *Drosophila P* transposons were mobile in zebrafish but have been domesticated in a common ancestor of chicken and human." Mol Biol Evol **22**(4): 833-44.

Haren, L., B. Ton-Hoang, et al. (1999). "Integrating DNA: transposases and retroviral integrases." Annu Rev Microbiol **53**: 245-81.

Hayashi, S., K. Ito, et al. (2002). "GETDB, a database compiling expression patterns and molecular locations of a collection of Gal4 enhancer traps." Genesis **34**(1-2): 58-61.

Hendrix, D. A., J. W. Hong, et al. (2008). "Promoter elements associated with RNA Pol II stalling in the *Drosophila* embryo." Proc Natl Acad Sci U S A **105**(22): 7762-7.

Hertz, G. Z. and G. D. Stormo (1999). "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences." Bioinformatics **15**(7-8): 563-77.

Hiromi, Y., A. Kuroiwa, et al. (1985). "Control elements of the *Drosophila* segmentation gene fushi tarazu." Cell **43**(3 Pt 2): 603-13.

Hochheimer, A., S. Zhou, et al. (2002). "TRF2 associates with DREF and directs promoter-selective gene expression in *Drosophila*." Nature **420**(6914): 439-45.

Hormozdiari, F., C. Alkan, et al. (2010). "*Alu* repeat discovery and characterization within human genomes." Genome Res.

Hoskins, R. A., J. M. Landolin, et al. (2010). "Genome-wide analysis of promoter architecture in *Drosophila melanogaster*." Genome Res.

Hu, W. Y. and K. M. Derbyshire (1998). "Target choice and orientation preference of the insertion sequence IS903." J Bacteriol **180**(12): 3039-48.

Hu, W. Y., W. Thompson, et al. (2001). "Anatomy of a preferred target site for the bacterial insertion sequence IS*903*." J Mol Biol **306**(3): 403-16.

Hutchison, C. A., S. N. Peterson, et al. (1999). "Global transposon mutagenesis and a minimal *Mycoplasma* genome." Science **286**(5447): 2165-9.

Isogai, Y., S. Keles, et al. (2007). "Transcription of histone gene cluster by differential core-promoter factors." Genes Dev **21**(22): 2936-49.

Ito, T., R. Motohashi, et al. (2005). "A resource of 5,814 dissociation transposon-tagged and sequence-indexed lines of *Arabidopsis* transposed from start loci on chromosome 5." Plant Cell Physiol **46**(7): 1149-53.

Janitz, M. (2008). Next-generation genome sequencing : towards personalized medicine. Weinheim, Wiley-VCH ; Chichester : John Wiley [distributor].

Julian, A. M. (2003). Use of Bioinformatics to investigate an analyze transposable element insertions in the genomes of *Caenorhabditis elegans and Drosophila melanogaster* and into the target plasmid PGDV1, Texas A&M University. **Master of Science:** 123.

Juven-Gershon, T., J. Y. Hsu, et al. (2006). "Perspectives on the RNA polymerase II core promoter." Biochem Soc Trans **34**(Pt 6): 1047-50.

Kadonaga, J. T. (2002). "The DPE, a core promoter element for transcription by RNA polymerase II." Exp Mol Med **34**(4): 259-64.

Kaminker, J. S., C. M. Bergman, et al. (2002). "The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective." Genome Biol **3**(12): RESEARCH0084.

Kapitonov, V. V. and J. Jurka (2003). "Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome." Proc Natl Acad Sci U S A **100**(11): 6569-74.
Karess, R. E. and G. M. Rubin (1984). "Analysis of P transposable element functions in Drosophila." Cell **38**(1): 135-46.
Kassis, J. A., E. Noll, et al. (1992). "Altering the insertional specificity of a Drosophila transposable element." Proc Natl Acad Sci U S A **89**(5): 1919-23.
Kassis, J. A., E. P. VanSickle, et al. (1991). "A fragment of engrailed regulatory DNA can mediate transvection of the white gene in *Drosophila*." Genetics **128**(4): 751-61.
Kaufman, P. D., R. F. Doll, et al. (1989). "*Drosophila* P element transposase recognizes internal P element DNA sequences." Cell **59**(2): 359-71.
Kaufman, P. D. and D. C. Rio (1992). "P element transposition in vitro proceeds by a cut-and-paste mechanism and uses GTP as a cofactor." Cell **69**(1): 27-39.
Kelley, M. R., S. Kidd, et al. (1987). "Restriction of P-element insertions at the Notch locus of *Drosophila melanogaster*." Mol Cell Biol **7**(4): 1545-8.
Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-64.
Kivistik, P. A., M. Kivisaar, et al. (2007). "Target site selection of *Pseudomonas putida* transposon Tn*4652*." J Bacteriol **189**(10): 3918-21.
Kornberg, T. B. and M. A. Krasnow (2000). "The *Drosophila* genome sequence: implications for biology and medicine." Science **287**(5461): 2218-20.
Korswagen, H. C., R. M. Durbin, et al. (1996). "Transposon *Tc1*-derived, sequence-tagged sites in *Caenorhabditis elegans* as markers for gene mapping." Proc Natl Acad Sci U S A **93**(25): 14680-5.
Kumar, A., M. Seringhaus, et al. (2004). "Large-scale mutagenesis of the yeast genome using a *Tn7*-derived multipurpose transposon." Genome Res **14**(10A): 1975-86.
Kuromori, T., T. Hirayama, et al. (2004). "A collection of 11 800 single-copy Ds transposon insertion lines in Arabidopsis." Plant J **37**(6): 897-905.
Kutach, A. K. and J. T. Kadonaga (2000). "The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters." Mol Cell Biol **20**(13): 4754-64.
Lee, C. C., E. L. Beall, et al. (1998). "DNA binding by the KP repressor protein inhibits P-element transposase activity in vitro." Embo J **17**(14): 4166-74.
Lee, C. C., Y. M. Mul, et al. (1996). "The *Drosophila* P-element KP repressor protein dimerizes and interacts with multiple sites on P-element DNA." Mol Cell Biol **16**(10): 5616-22.
Lee, I. and R. M. Harshey (2003). "Patterns of sequence conservation at termini of long terminal repeat (LTR) retrotransposons and DNA transposons in the human genome: lessons from phage Mu." Nucleic Acids Res **31**(15): 4531-40.
Lerman, D. N., P. Michalak, et al. (2003). "Modification of heat-shock gene expression in *Drosophila melanogaster* populations via transposable elements." Mol Biol Evol **20**(1): 135-44.
Liao, G. C., E. J. Rehm, et al. (2000). "Insertion site preferences of the P transposable element in *Drosophila melanogaster*." Proc Natl Acad Sci U S A **97**(7): 3347-51.
Lim, C. Y., B. Santoso, et al. (2004). "The MTE, a new core promoter element for transcription by RNA polymerase II." Genes Dev **18**(13): 1606-17.

Linheiro, R. S. and C. M. Bergman (2008). "Testing the palindromic target site model for DNA transposon insertion using the *Drosophila melanogaster* P-element." Nucleic Acids Res **36**(19): 6199-208.

Mackay, T. F. C., S. Richards, et al. (2008). "Proposal to sequence a *Drosophila* Genetic Reference Panel: a community resource for the study of genotypic and phenotypic variation. White Paper, NHGRI.", from http://www.hgsc.bcm.tmc.edu/project-species-i-Drosophila_genRefPanel.hgsc.

Mardis, E. R. (2008). "Next-generation DNA sequencing methods." Annu Rev Genomics Hum Genet **9**: 387-402.

Mates, L., Z. Izsvak, et al. (2007). "Technology transfer from worms and flies to vertebrates: transposition-based genome manipulations and their future perspectives." Genome Biol **8 Suppl 1**: S1.

Mavrich, T. N., C. Jiang, et al. (2008). "Nucleosome organization in the *Drosophila* genome." Nature **453**(7193): 358-62.

Medini, D., D. Serruto, et al. (2008). "Microbiology in the post-genomic era." Nat Rev Microbiol **6**(6): 419-30.

Merriman, P. J., C. D. Grimes, et al. (1995). "S elements: a family of Tc1-like transposons in the genome of *Drosophila melanogaster*." Genetics **141**(4): 1425-38.

Metaxakis, A., S. Oehler, et al. (2005). "*Minos* as a genetic and genomic tool in *Drosophila melanogaster*." Genetics **171**(2): 571-81.

Misra, S. and D. C. Rio (1990). "Cytotype control of *Drosophila* P element transposition: the 66 kd protein is a repressor of transposase activity." Cell **62**(2): 269-84.

Mori, I., G. M. Benian, et al. (1988). "Transposable element *Tc1* of *Caenorhabditis elegans* recognizes specific target sequences for integration." Proc Natl Acad Sci U S A **85**(3): 861-4.

Mul, Y. M. and D. C. Rio (1997). "Reprogramming the purine nucleotide cofactor requirement of *Drosophila* P element transposase in vivo." Embo J **16**(14): 4441-7.

Muller, F., M. A. Demeny, et al. (2007). "New problems in RNA polymerase II transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors." J Biol Chem **282**(20): 14685-9.

Mullins, M. C., D. C. Rio, et al. (1989). "cis-acting DNA sequence requirements for P-element transposition." Genes Dev **3**(5): 729-38.

Muse, G. W., D. A. Gilchrist, et al. (2007). "RNA polymerase is poised for activation across the genome." Nat Genet **39**(12): 1507-11.

Myers, E. W., G. G. Sutton, et al. (2000). "A whole-genome assembly of *Drosophila*." Science **287**(5461): 2196-204.

Nassif, N., J. Penney, et al. (1994). "Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair." Mol Cell Biol **14**(3): 1613-25.

O'Brochta, D. A., C. D. Stosic, et al. (2009). "Transpositionally active episomal hAT elements." BMC Mol Biol **10**: 108.

O'Brochta, D. A., W. D. Warren, et al. (1994). "Interplasmid transposition of *Drosophila hobo* elements in non-drosophilid insects." Mol Gen Genet **244**(1): 9-14.

O'Hare, K., A. Driver, et al. (1992). "Distribution and structure of cloned P elements from the *Drosophila melanogaster* P strain pi 2." Genet Res **60**(1): 33-41.

O'Hare, K. and G. M. Rubin (1983). "Structures of P transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome." Cell **34**(1): 25-35.
O'Kane, C. J. and W. J. Gehring (1987). "Detection in situ of genomic regulatory elements in *Drosophila*." Proc Natl Acad Sci U S A **84**(24): 9123-7.
Ohler, U. (2006). "Identification of core promoter modules in *Drosophila* and their application in accurate transcription start site prediction." Nucleic Acids Res **34**(20): 5943-50.
Ohler, U., G. C. Liao, et al. (2002). "Computational analysis of core promoters in the *Drosophila* genome." Genome Biol **3**(12): RESEARCH0087.
Ohtsuki, S., M. Levine, et al. (1998). "Different core promoters possess distinct regulatory activities in the *Drosophila* embryo." Genes Dev **12**(4): 547-56.
Pettersson, E., J. Lundeberg, et al. (2009). "Generations of sequencing technologies." Genomics **93**(2): 105-11.
Preclin, V., E. Martin, et al. (2003). "Target sequences of *Tc1*, *Tc3* and *Tc5* transposons of *Caenorhabditis elegans*." Genet Res **82**(2): 85-8.
Preston, C. R., J. A. Sved, et al. (1996). "Flanking duplications and deletions associated with P-induced male recombination in *Drosophila*." Genetics **144**(4): 1623-38.
R Development Core Team. (2008). R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.
R Development Core Team (2009). R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.
Rabenstein, M. D., S. Zhou, et al. (1999). "TATA box-binding protein (TBP)-related factor 2 (TRF2), a third member of the TBP family." Proc Natl Acad Sci U S A **96**(9): 4791-6.
Rach, E. A., H. Y. Yuan, et al. (2009). "Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the *Drosophila* genome." Genome Biol **10**(7): R73.
Reiss, D., H. Quesneville, et al. (2003). "Hoppel, a P-like element without introns: a P-element ancestral structure or a retrotranscription derivative?" Mol Biol Evol **20**(6): 869-79.
Ringrose, L. and R. Paro (2007). "Polycomb/Trithorax response elements and epigenetic memory of cell identity." Development **134**(2): 223-32.
Rio, D. C. (2002). P transposable elements in *Drosophila melanogaster*. Mobile DNA II. N. Craig. Washington, D.C., ASM Press**:** 484-518.
Rio, D. C., F. A. Laski, et al. (1986). "Identification and immunochemical analysis of biologically active Drosophila P element transposase." Cell **44**(1): 21-32.
Rio, D. C. and G. M. Rubin (1988). "Identification and purification of a Drosophila protein that binds to the terminal 31-base-pair inverted repeats of the P transposable element." Proc Natl Acad Sci U S A **85**(23): 8929-33.
Roiha, H., G. M. Rubin, et al. (1988). "P element insertions and rearrangements at the singed locus of Drosophila melanogaster." Genetics **119**(1): 75-83.
Rong, Y. S. and K. G. Golic (2000). "Gene targeting by homologous recombination in Drosophila." Science **288**(5473): 2013-8.
Ross-Macdonald, P., P. S. Coelho, et al. (1999). "Large-scale analysis of the yeast genome by transposon tagging and gene disruption." Nature **402**(6760): 413-8.
Rubin, G. M., M. G. Kidwell, et al. (1982). "The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations." Cell **29**(3): 987-94.
Rubin, G. M., M. D. Yandell, et al. (2000). "Comparative genomics of the eukaryotes." Science **287**(5461): 2204-15.

Ryder, E., F. Blows, et al. (2004). "The DrosDel collection: a set of P-element insertions for generating custom chromosomal aberrations in Drosophila melanogaster." Genetics **167**(2): 797-813.
Ryder, E. and S. Russell (2003). "Transposable elements as tools for genomics and genetics in Drosophila." Brief Funct Genomic Proteomic **2**(1): 57-71.
Sandelin, A., W. Alkema, et al. (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." Nucleic Acids Res **32**(Database issue): D91-4.
Schneider, T. D. and R. M. Stephens (1990). "Sequence logos: a new way to display consensus sequences." Nucleic Acids Res **18**(20): 6097-100.
Schuettengruber, B., D. Chourrout, et al. (2007). "Genome regulation by polycomb and trithorax proteins." Cell **128**(4): 735-45.
Schuettengruber, B., M. Ganapathi, et al. (2009). "Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos." PLoS Biol **7**(1): e13.
Schuster, S. C. (2008). "Next-generation sequencing transforms today's biology." Nat Methods **5**(1): 16-8.
Searles, L. L., A. L. Greenleaf, et al. (1986). "Sites of P element insertion and structures of P element deletions in the 5' region of Drosophila melanogaster RpII215." Mol Cell Biol **6**(10): 3312-9.
Seringhaus, M., A. Kumar, et al. (2006). "Genomic analysis of insertion behavior and target specificity of mini-Tn7 and Tn3 transposons in Saccharomyces *cerevisiae*." Nucleic Acids Res **34**(8): e57.
Shilova, V. Y., D. G. Garbuz, et al. (2006). "Remarkable site specificity of local transposition into the Hsp70 promoter of Drosophila melanogaster." Genetics **173**(2): 809-20.
Smale, S. T. and J. T. Kadonaga (2003). "The RNA polymerase II core promoter." Annu Rev Biochem **72**: 449-79.
Smit, A. F. (1999). "Interspersed repeats and other mementos of transposable elements in mammalian genomes." Curr Opin Genet Dev **9**(6): 657-63.
Smith, G. C. and S. P. Jackson (1999). "The DNA-dependent protein kinase." Genes Dev **13**(8): 916-34.
Spradling, A. C. and G. M. Rubin (1982). "Transposition of cloned P elements into Drosophila germ line chromosomes." Science **218**(4570): 341-7.
Spradling, A. C., D. Stern, et al. (1999). "The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes." Genetics **153**(1): 135-77.
Spradling, A. C., D. M. Stern, et al. (1995). "Gene disruptions using P transposable elements: an integral component of the Drosophila genome project." Proc Natl Acad Sci U S A **92**(24): 10824-30.
Staden, R. (1989). "Methods for calculating the probabilities of finding patterns in sequences." Comput Appl Biosci **5**(2): 89-96.
Stajich, J. E., D. Block, et al. (2002). "The Bioperl toolkit: Perl modules for the life sciences." Genome Res **12**(10): 1611-8.
Staveley, B. E., T. R. Heslip, et al. (1995). "Protected P-element termini suggest a role for inverted-repeat-binding protein in transposase-induced gap repair in *Drosophila melanogaste*r." Genetics **139**(3): 1321-9.
Taillebourg, E. and J. M. Dura (1999). "A novel mechanism for P element homing in Drosophila." Proc Natl Acad Sci U S A **96**(12): 6856-61.

Tang, M., C. Cecconi, et al. (2007). "Analysis of P element transposase protein-DNA interactions during the early stages of transposition." J Biol Chem **282**(39): 29002-12.
Tang, M., C. Cecconi, et al. (2005). "Guanosine triphosphate acts as a cofactor to promote assembly of initial P-element transposase-DNA synaptic complexes." Genes Dev **19**(12): 1422-5.
Tenzen, T. and E. Ohtsubo (1991). "Preferential transposition of an IS*630*-associated composite transposon to TA in the 5'-CTAG-3' sequence." J Bacteriol **173**(19): 6207-12.
Thibault, S. T., M. A. Singer, et al. (2004). "A complementary transposon tool kit for *Drosophila melanogaster* using *P* and *piggyBac*." Nat Genet **36**(3): 283-7.
Timakov, B., X. Liu, et al. (2002). "Timing and targeting of P-element local transposition in the male germline cells of Drosophila melanogaster." Genetics **160**(3): 1011-22.
Tower, J., G. H. Karpen, et al. (1993). "Preferential transposition of Drosophila P elements to nearby chromosomal sites." Genetics **133**(2): 347-59.
Tsubota, S., M. Ashburner, et al. (1985). "P-element-induced control mutations at the r gene of Drosophila melanogaster." Mol Cell Biol **5**(10): 2567-74.
Tudor, M., M. Lobocka, et al. (1992). "The pogo transposable element family of Drosophila melanogaster." Mol Gen Genet **232**(1): 126-34.
van Luenen, H. G., S. D. Colloms, et al. (1994). "The mechanism of transposition of Tc3 in *C. elegans*." Cell **79**(2): 293-301.
Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.
Vigdal, T. J., C. D. Kaufman, et al. (2002). "Common physical properties of DNA affecting target site selection of sleeping beauty and other Tc1/mariner transposable elements." J Mol Biol **323**(3): 441-52.
Viggiano, L., C. Caggese, et al. (1997). "Cloning and characterization of a copy of Tirant transposable element in *Drosophila melanogaster*." Gene **197**(1-2): 29-35.
Voelker, R. A., J. Graves, et al. (1990). "Mobile element insertions causing mutations in the Drosophila suppressor of sable locus occur in DNase I hypersensitive subregions of 5'-transcribed nontranslated sequences." Genetics **126**: 1071-82.
Walser, J. C., B. Chen, et al. (2006). "Heat-shock promoters: targets for evolution by P transposable elements in Drosophila." PLoS Genet **2**(10): e165.
Wang, W. C., F. M. Lin, et al. (2009). "miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression." BMC Bioinformatics **10**: 328.
Ward, E. J., I. Thaipisuttikul, et al. (2002). "GAL4 enhancer trap patterns during Drosophila development." Genesis **34**(1-2): 46-50.
Warren, W. C., L. W. Hillier, et al. (2008). "Genome analysis of the platypus reveals unique signatures of evolution." Nature **453**(7192): 175-83.
Weinert, B. T., B. Min, et al. (2005). "P element excision and repair by non-homologous end joining occurs in both G1 and G2 of the cell cycle." DNA Repair (Amst) **4**(2): 171-81.
Whalen, J. H. and T. A. Grigliatti (1998). "Molecular characterization of a retrotransposon in *Drosophila melanogaster*, *nomad*, and its relationship to other retrovirus-like mobile elements." Mol Gen Genet **260**(5): 401-9.
Witherspoon, D. J., J. Xing, et al. (2010). "Mobile element scanning (ME-Scan) by targeted high-throughput sequencing." BMC Genomics **11**: 410.

Wu, X., Y. Li, et al. (2005). "Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses." J Virol **79**(8): 5211-4.

Wysocka, J., T. Swigut, et al. (2006). "A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling." Nature **442**(7098): 86-90.

Yant, S. R., X. Wu, et al. (2005). "High-resolution genome-wide mapping of transposon integration in mammals." Mol Cell Biol **25**(6): 2085-94.

Zeitlinger, J., A. Stark, et al. (2007). "RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo." Nat Genet **39**(12): 1512-6.

Zhang, P. and A. C. Spradling (1993). "Efficient and dispersed local P element transposition from Drosophila females." Genetics **133**(2): 361-73.

Zhu, Q. and M. S. Halfon (2009). "Complex organizational structure of the genome revealed by genome-wide analysis of single and alternative promoters in *Drosophila melanogaster*." BMC Genomics **10**: 9.