Artificial horizontal transfer of retroposons.

Yeoh, Joseph Guan Chong

ARTIFICIAL HORIZONTAL TRANSFER OF RETROPOSONS

Thesis Submitted for the Degree of Doctor of Philosophy

Joseph **Yeoh** Guan Chong

School of Biological and Chemical Sciences

Queen Mary University of London

Statement of Originality

I, Joseph Yeoh, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signed:

Joseph Yeoh Guan Chong

13 October 2014

Acknowledgements

First and foremost, I want to thank my family for their continuous support, especially to my father. Without him, this PhD would not have been possible. He and my beloved mother have taught me valuable life lessons and I thank them for providing guidance, encouragement and support.

I thank Colin Malcolm and Richard Nichols for being excellent supervisors. I thank them for the time they have given to this project. I am particularly grateful for their advice and guidance in bioinformatics as well as laboratory techniques, but most importantly in developing and moulding my critical and decision making process, and hopefully I have made the transition from a student mind-set into a scientific worldview.

I am also thankful to the various people I have worked with throughout my PhD. I thank Andrew Leitch, who was the chair of my panel meetings, keeping my research on track. I also thank Taif Adams and Ruth Rose for lessons in molecular biology. I am grateful to Marleen Klann for her skills and guidance in germline transformation. Iskander Ibrahim and Stephen Rowden have given me advice on many parts of the research. Jasmin Zohren and Bruno Vieira assisted me with their IT and programming skills. I have learned a lot from the LabJolly Journal Club. Also, a special mention to the 5th floor Banter Lunch Group. The conversations we had have enriched my life.

I am grateful to Queen Mary for funding my project.

Abstract

Many factors may explain why certain transposable elements (TEs) spread in some species and not others. On the one hand, they include processes that affect the rate of transposition, such as differences in the regulation of expression; on the other hand, they include characteristics of a genome that affect the consequences of transposition. In particular genome size may have an effect: a genome that is large due to non-essential repetitive DNA may be permissive for TE movement, as insertion events are less likely to be deleterious. Genome size may also help explain the pattern of TE distribution between species of mosquitoes, including the important vectors of arboviruses, *Aedes aegypti* and *Culex pipiens* sensu lato. These species have genomes 3-5 times larger than a third genus, the *Anopheles* mosquitoes, which includes the malaria vectors. While all mosquitoes carry a diverse range of TEs, only culicines have the super abundant retroposon, Juan which can contribute up to 3% of the genome.

The genome sequences of various insect species were compared and the mosquitoes show a significant trend of increase in genome size, which can be attributed to the increase in retroposon sequences.

Two variants of Juan are reported, and new information is added regarding these elements. Previous publication of these elements contained errors in their sequences. A unique triple repeat of a cysteine rich region with a CCHC motif is present in the open reading frame. This sequence is a zinc-knuckle domain, important for the replication mechanism of these elements.

In comparison, a third recently active but very low copy number retroposon, termed Pip1, is also described. The results show that Pip1 is related to the Juan elements and also possess the triple CCHC motif. The PCR results also supports previous findings of polymorphism in insertion sites of this element, suggesting that Pip1 was active after the establishment of the different strains. Pip1 copies can be

grouped into three distinct groups based on nucleotide differences. Pip1 could also be using an alternative start codon to initiate transcription.

Full length intact copies of the three TEs in this study were been cloned into a germline transformation vector based on piggyBac and used for germline transformation in *Drosophila melanogaster. Drosophila melanogaster* has no Juan or Pip1 elements and an even smaller genome than anophelines mosquitoes, so insertion events from unregulated TE movement should be more detectable. We found that the elements have been successfully introduced into the *Drosophila* lines. The lines were inbred to obtain a homozygous population. A range of transformed lines were monitored. No effects of hybrid dysgenesis was found. Flies with black spotted eyes were identified in a Pip1 line but this phenotype was not heritable. Whole genome sequencing was carried out on the flies using next generation sequencing (NGS) technology. Retroposon sequences was detected at a high frequency. Insertion junctions were not detected but this result does not eliminate the possibility that a junction is present but the sequencing was not sensitive enough. A possible explanation is the retroposon is present as extrachromosomal plasmid DNA.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

aa	amino acids
BAC	Bacterial Artificial Chromosomes
BLAST	Basic Local Alignment Search Tool
EDTA	Ethylenediaminetetraacetic
IPTG	Isopropyl β-D-1-thiogalactopyranoside
LB broth	Lysogeny broth
LINE	Long Interspersed Nuclear Elements
LTR	Long terminal repeats
NaCl	Sodium Chloride
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
ORF	Open reading frame
PCR	Polymerase Chain Reaction
Rpm	Revolutions per minute
RT	Reverse transcriptase
SDS	Sodium dodecyl sulfate
TAE buffer	Tris base- acetic acid- EDTA buffer
TE	Transposable elements
TrisHCL	Tris(hydroxymethyl)aminomethane hydrochloric acid
TSD	Target site duplication
UV light	Ultraviolet light
X-GAL	Indoxyl 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside

CHAPTER 1

GENERAL INTRODUCTION

Johanes Gutenberg's printing machine was the beginning of the Renaissance and revolutionised the world. Books were produced quickly and *en masse*, thus spreading the ideas they carry to a much wider audience. This technology has now proceeded one step further. We now live in an era where it is now even possible to print an object in 3 dimensions! In fact, a 3D printer which can make most of its own components has been created (some assembly required). RepRap is a self-replicating manufacturing machine; the information required to produce it is open-sourced (RepRap, 2013). Perhaps appropriately, RepRap 1.0 and RepRap 2.0 are named Darwin and Mendel respectively, to reflect the replicative and evolutionary nature of this printer.

There is a broad analogy between such a self-replicating device and transposable elements (TEs). TEs are mobile genetic elements present in genomes. They are autonomous elements that encode domains that enable them to replicate themselves outside host DNA replication control (Wicker *et al*, 2007). However, just as the RepRap printers cannot replicate themselves independently of human society, TEs utilises the biochemical supplies of the cell to ensure its own replication. Despite these constraints, TEs can replicate to reach a very high copy number within their host genome.

In contrast, non-autonomous elements rely on the replication machinery of autonomous elements to mobilize themselves (Wicker *et al*, 2007). These elements hijack the proteins from autonomous elements to mobilise. Without the proteins coded by autonomous elements, non-autonomous elements would be unable to replicate. It is also possible to find 'relics' in many genomes; this term is used to

describe TE copies that have acquired mutations which render them unable to transpose, such as nonsense mutations to their ORFs. However, relics still share sufficient sequence identity with intact elements to allow their identification.

1.1 Transposable Elements

Since the initial discovery of TEs by McClintock (1956), TEs have been found in almost all genomes. TEs are divided into 2 broad classes: Class I and Class II (Wicker *et al*, 2007). TEs that mobilize via a DNA intermediate are grouped into Class II. They are also referred to as transposons in the literature. Transposons are distinguishable by their inverted terminal repeats at their ends as well as a single ORF. The ORF encodes a transposase which excises the element from the genetic loci and inserts it elsewhere in the genome. This mechanism is often referred to as cut and paste. Examples of transposons are piggyBac (Cary *et al*, 1989) and P elements (Engels, 1989).

Class I elements are elements that mobilize through an RNA intermediate. Class I elements are further divided into 2 different categories based on their overall structure. Elements which have terminal repeats at their ends are referred to as retrotransposons or long terminal repeat (LTR) retrotransposons. The LTRs are important for replication initiation and termination of transcription. Retrotransposons also contain open reading frames (ORFs) similar to retroviral products such as *gag*, RNase H and integrase (Eickbush and Malik, 2002). An example is the gypsy retrotransposon (Kim *et al*, 1994).

The other category of Class I elements are the retroposons. However, these elements do not contain LTR; thus, these elements are also described as non-LTR-retrotransposons or long interspersed nuclear elements (LINE) in the literature. Most retroposons contain ORFs which encode for reverse transcriptase, endonuclease and a nucleic acid binding domain (Malik et al, 1999). Examples of retroposons include

LINE-1 elements (Moran and Gilbert, 2002) and Juan elements (Mouches *et al*, 1992; Agarwal *et al*, 1993).

TEs are important components of genomes. Their ability to replicate outside host control has allowed them to achieve high copy numbers within the host genome. This thesis focuses on specific Class I elements (the retroposons), and their relationship to their host genome.

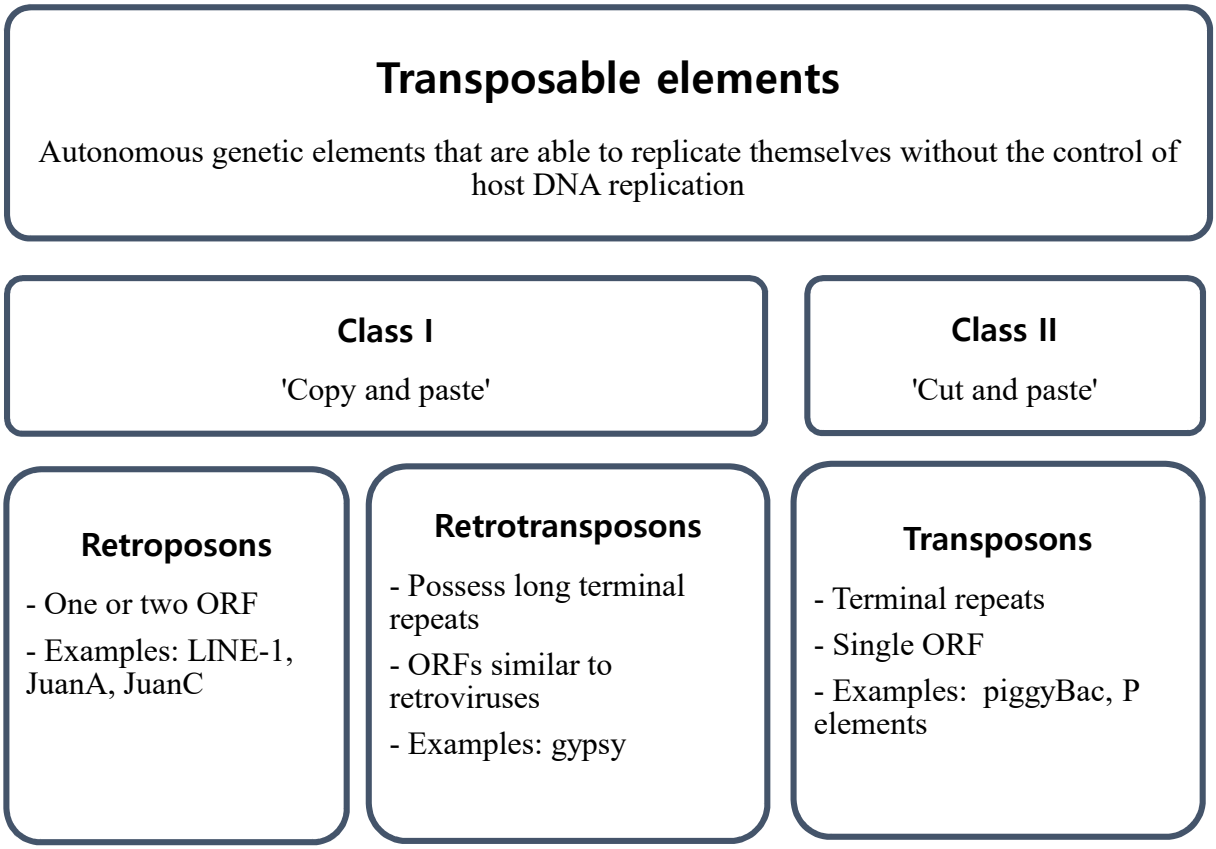

Fig 1.1 Classification of eukaryotic transposable elements (TEs). Class I elements uses an mRNA as an intermediate while in Class II (the transposons), there is no mRNA intermediate.

1.2 Retroposons

All autonomous retroposons contain the coding domain for the enzyme reverse transcriptase and the endonuclease domain. The reverse transcriptase enzyme catalyzes the reverse transcription of mRNA into cDNA while the endonuclease generates the nick at the target site (Han, 2010). The simplest retroposon, R2, contains only these coding domains. Other types of retroposons have an additional coding domain- the nucleic acid binding domain. This domain may play a role in binding the protein to the retroposon mRNA (Eickbush and Malik, 2002).

Retroposons do not contain long terminal repeats; hence they are sometimes named non-LTR retrotransposons. However, they contain an A-rich 3' tail. It is unfortunate that retroposons are more often defined for a structure that they do not possess rather than something that they do possess. Another unhelpful term is 'LINEs'; which derives from the original description of Long Interspersed Nuclear Elements in the human genome. However, these names does not distinguish retroposons from other repetitive elements, further confusing many undergraduate students when they first venture into the wealth of mobile element literature, me included. Thus, Eickbush and Malik (2002) have repeated the need to call these elements 'retroposons', a term which refers to the nature of their replication process.

Retroposons replicate via an RNA intermediate akin to a 'copy-and-paste' function of a text editor. The process is termed target primed reverse transcription (Christensen & Eickbush, 2005; Han, 2010). Firstly, an autonomous element is transcribed and the mRNA exported from the nucleus. Translation of the ORFs produce retroposon proteins, and these proteins bind to the mRNA, forming a ribonucleoprotein complex. This complex is transported back into the nucleus. One of the strands of DNA is cleaved, forming the potential target insertion site. The cleavage is carried out by the endonuclease. Minus strand synthesis is carried out by the reverse transcriptase using the mRNA as the template. At some point during or after the

strand synthesis, the other DNA strand is also cleaved. Using the newly synthesised retroposon DNA, the plus strand is synthesised. After synthesis is completed, the mRNA template is removed and any gaps filled by host cell proteins. This process generates target site duplications and is a hallmark of target primed reverse transcription. However, the reverse transcription process can terminate before the whole sequence is reverse transcribed, generating copies with 5′ truncations. Retroposons contain poly-A tails at the 3′ end because the mRNA is reverse transcribed.

There is no evidence of retroposons undergoing horizontal transfer. Horizontal transfer is the transmission of DNA information from one organism to another organism. Thus, a pre-requirement for horizontal transmission is an infective particle capable of moving successfully from host genome into another genome of another species. The only infective particle during the retroposon life cycle is the mRNA generated. RNA is less stable than DNA and therefore would be degraded quickly when it leaves a cellular environment. Unlike an RNA virus, retroposons cannot package the RNA into a protective protein coat. Therefore, due to the nature of their replication process retroposons do not have effective vectors for horizontal transmission (Eickbush and Malik, 2002). In fact, there are some reports of retroposons having undergone horizontal transfer (Mouches et al, 1992), but these were later discredited by the collection of more data (Biedler and Tu, 2007).

Scouring the literature, there has only been one publication of a successful introduction of a retroposon in insects in the laboratory. Eickbush *et al* (2000) introduced R2 sequences from the silkworm *Bombyx mori* into *Drosophila melanogaster*. They introduced purified R2 proteins and mRNA from *B. mori* and injected this mix into *D. melanogaster* embryos. By analysing the transformed flies, they found the *B. mori* R2 sequences at the 28S rRNA genes. However, R2 elements occur naturally in *D. melanogaster*. Thus, it is not surprising that R2 elements from the silkworm could integrate into the fruitfly genome.

Additionally, retroposons undergo strong purifying selection (Malik and Eickbush, 1999; Biedler and Tu, 2007). Evidence of this selection can be seen in the significantly higher rate of synonymous substitution compared to non-synonymous substitutions. As a retroposon can only replicate via a copy and paste mechanism, any copies which contain silencing mutations would not be able to replicate themselves and would be an evolutionary dead-end, eventually forming relics. In contrast, copies with intact reading frames will be able to replicate and persist over the generations.

Therefore, retroposons are under negative selection pressure. It would be more advantageous for the host to have inactivated copies of retroposons. Since retroposons are only transmitted vertically, a retroposon which continuously replicates and produces active copies would have a higher chance of remaining active in the host genome. This thesis explores the possibility of introducing retroposons from one species into another completely different species, especially retroposons that are not present in the recipient genome. This would provide answers as to the efficiency of horizontal transfer of retroposons if a mechanism allowing it to do so exists.

1.3 Transposable Element Activity in Genomes

For a long period of time, the prevailing view of genetics is that every bit of DNA in the genome must have a direct benefit to the host. Therefore, it was not too surprising when Barbara McClintock first published her results on Activator and Dissociation system in maize (McClintock, 1951), it was met with intense scepticism. However, as scientists began to study genetics in greater detail, more and more of these elements were uncovered, leading to the baffling question of what benefit these elements have for the genome. After all, if these elements are harmful or of zero advantage, why are they present in the genome at all? On the other hand, if they are beneficial, what are the advantages they confer since their coding domains do not encode any functional protein to the host?

To address this subject, it is important to consider TE activity and the potential effects it has on the host, both directly and indirectly. An element can affect host genes directly as a consequence of mobilization. By inserting into exons of host genes, the element disrupts host gene functions (Kidwell and Lisch, 2002). Transposable element mobilization into introns also affects host gene function. Splicing of the intron is affected and proteins with the correct amino acids are not produced. An example is a retrotransposon insertion into an intron in mice, causing cataracts during development (Talamas *et al*, 2006). TE insertion into non-coding regions also disrupt gene function by altering host gene regulation. Insertion of a Mu element into the intron of the knotted gene in maize prevents repression and causes ectopic expression in leaves (Kidwell and Lisch, 2002). In short, the insertion of transposable elements into exons, introns and non-coding regions can change expression of the host gene.

In addition, there is evidence that insertions near a gene can disrupt nearby gene function. One possible explanation for this effect is the action of silencing mechanisms to suppress the TE through chromatin modifications. For example, Slotkin and Martienssen (2007) found histone tail modification and DNA methylation on chromatin containing the TE, which are associated with changes in chromatin packing and condensation, ultimately forming a dense, packed, transcriptionally silenced heterochromatin in mouse embryonic stem cells. The methylation of histone 3 or methylation of cytosine residues in CpG islands (region with a high frequency of cytosine followed immediately by a guanine) is thought to act as a signal for formation of heterochromatin in mammals. We know most about pathways that could link heterochromatin formation with TE downregulation from studies on model organisms. For example, in *Arabidopsis*, the DDM1 gene coding for a *de novo* methyltransferase is important for maintaining DNA methylation patterns as well as methylation of histones, since in DDM1 mutants, DNA methylation is lost and histone 3 is not methylated (Gendrel *et al*, 2002). Upregulation of retrotransposons and transposons were detected in these mutants. A similar result was found in

experiments on mouse embryonic stem cells. H3 lysine 9 (H3K9) Suv39h histone methyltransferase gene is responsible for methylation of lysine 9 on histone 3. Elevated levels of TE transcripts were detected in cells deficient in this gene (Martens *et al*, 2005).

The pathway of TE silencing through chromatin modifications has been investigated by several groups (Lippman *et al*, 2004; Vagin *et al*, 2006; Slotkin and Martienssen, 2007). In summary, they infer that double-stranded RNA (dsRNA) transcribed from TE are cleaved into siRNA. These siRNAs form a complex with argonaute-family protein. The siRNA-protein complex then targets and cleaves RNA which are still being transcribed and attached to both an RNA polymerase II and the DNA strand. The cleaved RNA acts as a signal for other proteins, such as H3K9 methyltransferase, *de novo* DNA methyltransferase and other chromatin-modifying proteins. This leads to methylation of the histone or methylation of the cytosine bases, resulting in formation of heterochromatin. For example, in *C. elegans*, Argonaute proteins form a complex with small interfering RNAs (siRNA) that guide RNA-degrading complexes to complementary transcripts (Sijen and Plasterk, 2003). Evidence for a role in TE suppression was obtained from transgenic lines which are Argonaute deficient. These lines produced the siRNAs but could not target complementary transcripts. The siRNA of the transposon Tc1 and elevated transcripts of Tc1 were detected in the mutant lines.

There is evidence for similar pathways acting in insects, which are therefore likely to affect the mosquitoes in this study. In *Drosophila*, the gene Piwi encodes a protein of the Argonaute family. Northern blotting shows that Piwi protein associates with siRNA, and so may have an Argonaute-like function (Saito *et al*, 2006). Genes which are responsible for methylation of histones have also been identified in *Drosophila*. One of them, Enhancer of Zeste, produces a protein that shows methyltransferase activity that methylates lysine 9 and lysine 27 residues in histone 3 (Czermin *et al*, 2002), and these methylation marks are found in *Drosophila*

heterochromatin. This evidence suggests that at least one mode of TE suppression in insects could be the formation of the dense, transcriptionally silent heterochromatin.

To reduce the potential negative effects caused by insertions, some mobile elements have acquired the ability to insert specifically into a target region. The best example of this is the R1 and R2 elements in insects, which insert specifically into the 28S of ribosomal DNA. The endonuclease of the R2 element is very accurate and targets the DNA sequence (Eickbush, 2002). However, since rDNA exists in multiple tandem copies within the genome, insertions of these elements are not too detrimental. Another example is the Het-A and Tahre and TART elements in *D. melanogaster* (Mason *et al*, 2007). These elements mobilise after DNA replication and insert into the ends of chromosomes, forming the telomeres.

TEs compete with each other for metabolic resources. This includes the components needed for mobilisation, insertion sites in the genome and more importantly, the cost of reduced host survivability due to increased copy numbers (Leonardo and Nuzhdin, 2002). If more TEs are present within a host genome, the chances of host survivability is reduced. Thus, a TE will have better chances of survival if other mobile elements are not present.

TEs have also been found which have inserted into another TE, forming nested transposable elements (Kaminker *et al*, 2002; Weber and Schmidt, 2009). This has the twofold advantage of minimising damage to the host as well as inactivating another potential rival element. When a TE mobilises, it might insert into a gene sequence and have a negative impact on the host. However, if it inserts into another TE, it is unlikely to have a deleterious impact on the host as the resident TE would not be important for host survival. Moreover, the insertion will inactivate the resident element. The mobilising element has successfully generated a copy of itself as well as reducing active copies of other elements (Leonardo and Nuzhdin, 2002).

Another potential impact of TE activity is exon shuffling. When the element is mobilised, flanking sequences around the element is also copied and moved together with the element. If an exon is adjacent to the TE, this exon is copied and mobilised into a new region, potentially creating new protein products (Tautz and Domazet-Loso, 2011).

The enzymes involved in mobilization could also generate pseudogenes (Kidwell & Lisch, 2001; Lahn *et al*, 2001). Pseudogenes are DNA sequences which share sequence similarities to genes but do not encode for proteins or are not expressed. These pseudogenes do not have introns and because they are reverse transcribed from the mRNA, they also do not possess regulatory sequences such as promoters. The mRNA from a gene could be used by Class I mobilization machinery as a template and reverse transcribed into the genome.

Due to their abundance, some TEs have been recruited for host function. Exaptation is the process of a TE being adapted for host function. TE derived sequences form transcription factor binding sites, promoters, enhancers and silencers; thus, aiding in host gene regulation (Kidwell and Lisch, 2002; Guio *et al*, 2014). A more extreme example is the *D. melanogaster* telomeres (Mason *et al*, 2007). *D. melanogaster* has lost the telomerase enzyme and its function has been replaced by these elements. Het-A, Tahre and TART elements transpose from sites near the telomere ends to the ends of each chromosome after DNA replication.

TE activity also has an effect on a genomic level, by indirectly shaping the host genome. The genome of most organisms are organised into either a short or a long interspersion pattern. The *Aedes aegypti* and *Culex quinquefasciatus* genome has a short interspersion pattern: single copy genic sequences (<2000bp) are interrupted by non-genic sequences, including transposable elements (Rai, 2010). In contrast, *D. melanogaster* and the *Anopheles* genomes exhibit a long interspersion pattern: mobile

elements are present between unique, long, uninterrupted stretches of genic sequences (>13 000bp).

Transposable elements are also more commonly found away from gene-rich areas. *D. melanogaster* Het-A and Tahre and TART insert into the ends of chromosomes and form the telomeres (Mason *et al*, 2007). Other TEs are present at heterochromatin areas, or rather the high amount of TE at a particular region causes the formation of heterochromatin at that region. The human Y chromosome is abundant with TEs and insertion of TEs into the neo-Y gene causes heterochromatin formation and further reduced recombination with the neo-X gene (Kidwell and Lisch, 2002). The accumulation of TEs in X chromosomes has even facilitated the silencing of the extra X chromosome in human females: TE-rich areas are silenced and form heterochromatin first, and this signal slowly spreads along the chromosome (Slotkin and Martienssen, 2007).

An increase in copy number of an element has other indirect effects on the genome. The risk of non-homologous recombination and inversions increases as more of the genome consists of similar sequences. Non-homologous recombination occurs when two chromatids do not pair up equally, causing one strand to gain an extra stretch of DNA sequences while the other loses some DNA information. Ectopic recombination can also occur within a chromosome if there is high sequence similarity in the chromosome; for example, between two copies of a transposable element (Kidwell and Lisch, 2002). Schwartz *et al* (1998) found evidence for a LINE-mediated inversion event on one of the arms in the Y chromosome after the divergence of hominid and chimp lineages, but before the radiation of human populations.

The C-value paradox is another indirect effect of TE mobilisation in the genome. The C-value paradox refers is the very poor correlation between an organism's complexity and its genome size. Organisms that are more complex do not necessarily have a bigger genome and vice versa (Patrushev and Minkevich, 2008. For

example, salamanders have the biggest vertebrate genome (120Gbp) while humans have a genome size of 3.3Gbp). One of the factors causing this is the fact that TE mobilisation leads to an increase in DNA content in the genome without adding complexity to the host (Kidwell and Lisch, 2002). The total genomic DNA increases but the number of coding genes does not increase significantly. This is evident when genome sizes of different but closely related species are compared. The sequencing of the mosquito genomes provide an illustration of this. The dengue fever mosquito, *Aedes aegypti* has a genome size of 1.3Gbp, 15 419 genes and TE composition of 50% (Nene *et al*, 2007), while the genome of another mosquito, the West Nile Virus vector, *Culex quinquefasciatus*, has a genome size of 579Mbp, 18 883 genes and TE composition of 29% (Arensburger *et al*, 2010). The malarial mosquito, *Anopheles gambiae*, pales in comparison with a genome size of 278Mbp, 12 457 genes and 15% TE composition (Holt *et al*, 2002). The recently completed genome of *An. darlingi* (201Mbp) has a slightly smaller genome size to *An. gambiae* and 10 457 genes but a greatly reduced composition of TEs (2.29%) (Marinotti *et al*, 2013). The genome size increased by two-fold when comparing *Anopheles* with *Culex* and five-fold between *Anopheles* with *Aedes*, but the number of genes are in the range of 10 000 to 19 000 genes. There is poor correlation between mosquito genome size and complexity.

TE activity has such a huge impact on genomes that it has been speculated that their activity can drive the formation of new species (Kidwell and Lisch, 2002). Kidwell and Lisch's case is based on comparative data from a few sequenced genome species and their estimates of TE content. They cite the data obtained from studies on the bat genus *Myotis*. This genus has the most species of bats (103 species), and the genome of *M. lucifugus* contains a high number of TEs (Oliver and Greene, 2009). The TEs are still active and appear to amplify sporadically. They argue that the transposition would have created extra genetic variation, allowing adaptation to new environments. Whilst it is reasonable to propose a link between TE mobilisation and the creation of new genetic variation, the simultaneous occurrence of one single case

of adaptive radiation with a burst of TE activity is not particularly convincing evidence of a causal link. This could simply be a coincidence. There have been many proposed explanations for adaptive radiation and do not require a burst of mutations by TE activity. Speciation could be driven by other effects, including selection, founder effects, adaptive radiation due to changes in environment and reproductive isolation. None of these processes necessarily require a burst of TE activity in the genome before adaptive radiation occurs. For example, perhaps the most well-known example, the evolution of Darwin's finches in the Galapagos Islands, has been attributed to a combination of founder effects and selection imposed by food availability (Lamichhaney *et al*, 2015). Similarly, one of the most spectacular known radiations involves the Hawaiian *Drosophila*, a genus in which an estimated 1000 species are thought to have evolved since the Hawaiian Islands emerged in the last 25 million years. This group have radiated with changes in mating behaviour, feeding behaviour and geographical distribution, which have been explained by founder events, adaptation to newly arising environments and vicariance events (O'Grady *et al*, 2011). Such explanations do not require an elevated mutational input from TE activity, although it might accelerate the adaptation. Hence, in light of Oliver and Green's (2009) proposal it would be intriguing to investigate these and comparable radiations in a broader range of taxa to see if TE activity coincides more widely with adaptive radiation.

All of the effects described above are examples of TE behaviour within their host genome, but what happens when these elements cross into the genomes previously devoid of these elements? An example are the P and I element mobilisation in *D. melanogaster* (Rio, 2002; Spradling *et al*, 1999). The mobilisation of these elements gave rise to a phenomenon called 'hybrid dysgenesis'. Crosses between strains containing these elements with strains devoid of these elements produced offspring which suffered from various phenotypic abnormalities, such as reduced fecundity,

increased mutation rates and chromosomal rearrangements. Various mutations that affect the phenotype are observed due to these elements inserting into host genes.

In short, TE activity can affect host genes, shape genome organisation and ultimately shape the species. This thesis explores the activity of retroposons when they are introduced into a naïve genome.

1.4 Mosquito Genomes

A remarkable feature of mosquito genomes is the conservation of chromosome number. In the 300 mosquito species surveyed, all except one possesses 6 chromosomes, that is, 2n=6 (Rai, 2010). Mosquitoes are thought to have already evolved by 210 MYA. Despite the ancient origin of mosquitoes, speciation, chromosome repatterning and the emergence of sex chromosomes, the basic number of chromosomes has remained unchanged through all these processes.

Anopheles mosquitoes possess heteromorphic sex chromosomes while *Aedes* and *Culex* possess homomorphic sex chromosomes (sex is determined by a gene at a single locus) (Rai, 2010). There are two competing views for the evolution of sex chromosomes in the mosquitoes. The first hypothesis of sex chromosome evolution in mosquitoes is the view that heteromorphic sex chromosomes evolved from identical homologs. *Anopheles* mosquitoes evolved heteromorphic sex chromosomes while *Aedes* and *Culex* retained homomorphic sex chromosomes. Alternatively, if heteromorphic sex chromosomes are ancestral, then *Anopheles* retained the sex chromosomes while *Aedes* and *Culex* lost the smaller male-determining chromosome.

Another pattern also emerged from studying the genome size. Mosquitoes in the genus *Anopheles* have a genome size range of 0.23-0.29 pg/haploid genome (Rai, 2010). *Culex* species have a size range of 0.54-1.02 pg while *Aedes* species has the widest

range from 0.59 to 1.9 pg. Considering the fact that *Anopheles* are basal in the mosquito evolutionary tree, the genome sizes have shown an increasing trend.

Thirdly, the genome organization is also different between the *Anopheles* genomes and the other two genus (Tu and Coates, 2004). The *Cx. quinquefasciatus* and *Ae. aegypti* genomes are organised in a short-period interspersion pattern. Unique gene sequences, roughly 1-2 kbp in length are separated from each other by repetitive sequences not more than 4 kbp in length. The Anopheline genome shows a long-period interspersion pattern. Long stretches of unique gene sequences (>13 kbp) are interrupted by long (>5.6 kbp) repetitive elements.

The sequencing of four mosquito genomes and the deposition of this information into publicly available databases has also helped to further research in these mosquitoes. The first mosquito genome sequenced was *Anopheles gambiae*, (Holt *et al*, 2002), followed by *Aedes aegypti* (Nene *et al*, 2007), *Culex quinquefasciatus* (Arensburger *et al*, 2010) and lastly *Anopheles darlingi* (Marinotti *et al*, 2013). The smallest genome in the group is *An. darlingi* (201 Mbp), followed by *An. gambiae* (278 Mbp), *Cx. quinquefasciatus* (579 Mbp) and the largest genome sequenced is *Ae. aegypti* (1.3 Gbp). The amount of transposable elements in the genome also increases from *An. darlingi* (2.29%) to *An. gambiae* (15%) to *Cx. quinquefasciatus* (29%) and *Ae. aegypti* (50%). In addition, the genomes of these mosquitoes contain all classes of TEs described earlier in this chapter (Tu and Coates, 2004). They contain numerous novel elements that are not present in other species, such as the Juan elements (Mouches *et al*, 1992; Agarwal *et al*, 1993).

Their close relatedness to *Drosophila* has also aided the studies on the mosquitoes. *Drosophila melanogaster* last shared a common ancestor with the mosquitoes roughly 250 MYA. The genome of *D. melanogaster* has also been sequenced (Adams *et al*, 2000), and the 168 Mbp genome has 5.5% of transposable elements. Tools

and techniques developed for *D. melanogaster* can be used for the mosquitoes, such as germline transformation.

The ease of care of keeping mosquitoes in the laboratory as well as the medical importance in studying them has made them a model organism for studying insect vectors. With increasing fears that mosquito insecticide resistance is on the rise (Blair *et al*, 2000; Kyle and Harris, 2008; Edi *et al*, 2012), researchers have sought to control mosquito populations as an alternative to finding cures and treatments for diseases (Sinkins and Gould, 2006; Marshall and Taylor, 2009). Thus, the need to study mosquito genomes is ever greater than before.

1.5 Juan Elements and Pip1

The discovery of the Juan elements started with a study on insecticide resistance in *Culex pipiens* s.l. (Raymond *et al*, 1989). One mechanism of organophosphate insecticide resistance is an overabundance of non-specific esterases A and B, which can result from changes in gene expression or amplification of blocks of DNA containing one or two esterase genes. Amongst cases of the latter mechanism was a strain of *Culex quinquefasciatus* from California in which selection with insecticide had produced a very high copy number of the esterase B1 gene. Sequence analysis of the DNA co-amplified with the esterase gene led to the discovery of a disrupted Juan element, which in turn led to the discovery that it was a very abundant dispersed element in the genome (Mouches *et al*, 1992). The possibility that this element, which was termed JuanC, could have been involved in initiating the duplication and subsequent amplification of esterase genes prompted further investigations, which extended to *Aedes aegypti*.

Mouches *et al*, (1992) constructed a random genetic library of *Ae. aegypti* from the Pacific strain. The library was screened with an internal probe generated from the study on JuanC (Raymond *et al*, 1989). Recombinant phages which hybridised with

this probe were isolated. Two different cloned copies were entirely sequenced. The sequences contained only three differences and one of them, designated as JuanA1, was deposited into the NCBI database as the JuanA element (accession number M95171). The authors also estimated that 200 full length copies are present in the *Ae. aegypti* genome. In addition, the high degree of similarity of JuanA elements suggests a recent origin and amplification of these elements.

At this point, the full length sequence of JuanA was obtained but JuanC had not been completely sequenced. Consequently, Agarwal *et al* (1993) prepared a genomic library genomic library of *Culex pipiens* TEM-R strain to isolate JuanC elements. This library was screened using a fragment from the 5' end of JuanA. Recombinant phages which gave a positive signal to the probe were isolated, and two of the JuanC copies were entirely sequenced. The two sequences differed from each other due to substitutions or insertions at the 3' end. One of them was designated as JuanC1 and deposited into the NCBI database (accession number M91082). The estimated copy number for full length JuanC was 2500 copies per haploid genome, based on a genome size of 750 Mbp. Justas in the case of JuanA, the high degree of similarity present in JuanC copies suggests a recent origin and amplification of these elements.

Another subsequent paper was published on JuanA (Biedler and Tu, 2007). They used a combined bioinformatics plus PCR approach to characterise the element. The authors suggest that the contribution of JuanA to the *Aedes aegypti* genome is approximately 3%. It was also estimated that at least 378 copies of JuanA share 99% identity to the M95171 sequence. This is significant because these copies are potentially active and such a high number of potentially active copies are not normally found in a genome. They also identified Juan elements in other mosquito species, except for *Anopheles*. The significance of this paper was the usage of bioinformatics, together with PCR screens, to determine the presence of Juan elements in other

mosquito species. A significant finding was no Juan elements were detected in any *Anopheles* mosquitoes.

As initially described, both JuanA and JuanC are closely related and these elements share a few peculiar characteristics. Both elements are retroposons- they transpose via a RNA intermediate. Both are roughly 4.5kbp long and share high sequence identities with each other. Both elements are present in high copy numbers in their respective genomes. Both also show evidence of recent activity. The phylogenetic construction using sequences from open reading frame 2 (ORF2) also places them within the Jockey group of retroposon elements (Crainey *et al*, 2005).

1.6 Germline Transformation of Mosquitoes and Fruitflies

Germline transformation refers to the practice of microinjecting foreign DNA into an embryo in the early stages, and for the DNA to become incorporated into the germline. Therefore, the foreign DNA will be stably inherited in the offspring and subsequent generations. Among the insects, this method was first pioneered in the fruitfly, *Drosophila melanogaster*, by using P elements (Rubin and Spradling, 1982). This technology provided an excellent method to study *D. melanogaster* genetics and propelled *D. melanogaster* as a model organism for genetics. Moreover, this method provided an excellent tool to generate transgenic insects to control agricultural and medically important pests (Wimmer, 2003).

However, hopes that P elements could be used to transform other insects were not realised. P elements do not mobilise in other insect species (O'Brochta and Handler, 1988) and alternative methods was sought to enable germline transformation in other insects, particularly the mosquitoes. This section focuses on germline transformation methods used on the fruitfly and the mosquitoes.

The Hermes element was isolated from the housefly, *Musca domestica* (Warren *et al*, 1994). It is a transposon, thus mobilising using a cut-and-paste mechanism. Hermes was successfully used to transform *D. melanogaster*, *Ae. aegypti*. and *Cx.*

quinquefasciatus. The remobilization rates in the transformed insects were also measured. Remobilization of the introduced element could be beneficial because when it mobilises to a new gene location, it might alter the genes around that area, generating a phenotypic library, such as the gene disruption project using P elements in *D. melanogaster* (Spradling *et al*, 1999). On the other hand, it might not be beneficial to have the introduced gene remobilising in the transformed insect. If it mobilises into important genes, the mutation might be lethal and the transformed strain might be lost. For the Hermes element, remobilisation was detected in *D. melanogaster* but not in the mosquitoes (O'Brochta *et al*, 2003).

MosI mariner is another transposon which was used in germline transformation. First identified in *D. mauritiana*, the element was found in most other insect species and quite likely, this is due to horizontal transfer (Maruyama and Hartl, 1991; Robertson, 1993). It has been used to transform *D. melanogaster* and *Ae. aegypti* (Lidholm *et al*, 1993). Mariner does not remobilise in *D. melanogaster* but in *Ae. aegypti*, not only does it remobilise but it preferentially inserts into itself (O'Brochta *et al*, 2003).

Another element used for germline transformation of the fruitfly and mosquitoes is Minos. It is another transposon, isolated from another *Drosophila* species, *D. hydei* (Loukeris *et al*, 1995). It has been used to transform *D. melanogaster*, *An. stephensi* and *Ae. aegypti* (Catteruccia *et al*, 2000; O'Brochta *et al*, 2003; Metaxakis *et al*, 2005). Minos does not show evidence of remobilisation in any of the insects, although data on *D. melanogaster* is inconclusive (O'Brochta *et al*, 2003).

A fourth element, piggyBac, has also been used in germline transformation. This transposon is 2.4kbp, and was isolated from the cabbage looper moth, *Trichoplusia ni* (Cary *et al*, 1989; Fraser *et al*, 1995). It has been widely used to transform *D. melanogaster*, *An. gambiae* and *Ae. aegypti* (Grossman *et al*, 2001; Handler and Harrell, 2001; Lobo *et al*, 2002; Handler, 2002). In addition, transformation efficiency is quite high, reaching a high of 40% when transforming *An. albimanus*. Various markers have

also been added to the piggyBac vector to aid in the germline transformation process. piggyBac does not remobilise in the mosquitoes but there is evidence of remobilisation in *D. melanogaster* (O'Brochta *et al*, 2003).

A summary of the transposable elements used and tested is presented in Table 1.1, listing the species used and the remobilisation efficiency. Most of the elements have been tested and used in mosquito and fruitflies. However, all of these are Class II elements. Their rate of transposition also varies- they might not increase to a sufficient copy number before being inactivated. Moreover, transposons have the potential for horizontal transfer, thus there is a risk of spreading into non-target species. Therefore, there is still a need to identify elements that can be used for germline transformation (Sinkins and Gould, 2006).

Retroposons have the potential to be a useful germline transformation tool. However, their use for this purpose has largely remained unexplored. As mentioned earlier in the chapter, retroposons do not seem to undergo horizontal transfer, restricting their effects to their hosts. The Juan elements have also shown to be present in high copy number in their host genome- thus, they have the potential to spread and increase very quickly. The potential of using retroposons as insect germline transformation tools is therefore explored in this thesis.

Table 1.1 Transposon vector used in the germline transformation of fruitlfies and mosquitoes. Their lengths and mobilisation potential is also listed. References are available in the text.

	Length	*Drosophila*	*Anopheles*	*Aedes*	*Culex*	Remobilisation
Hermes	2.7kb	Yes	No evidence in literature	Yes	Yes	Yes in *Drosophila* None in *Aedes*
Mariner	1.3kb	Yes	No evidence in literature	Yes	No evidence in literature	Very low in *Drosophila* Yes in Aedes (inserts into itself)
Minos	1.8kb	Yes	Yes	Yes	No evidence in literature	None in all insects
piggyBac	2.5kb	Yes	Yes	Yes	No evidence in literature	Yes in *Drosophila* None in mosquitoes

1.7 Thesis outline

The main focus of the thesis is transposable elements in the mosquito genomes, particularly on retroposons). I explore why mosquito genomes differ, both in size and type of TE content. I also describe a few elements and characterise these retroposons. Finally, I explore their feasibility as germline transformation tools.

Chapter 2: General Materials and Methods

This chapter covers the various molecular biology methods and techniques used in the experiments. A more detailed materials and methods section is present in each chapter.

Chapter 3: Retroposons and the Mosquito Genome

In this chapter, I analyse the mosquito genomic sequencing data available and address the question of why their genome sizes and content differ. By examining their genomic content, I identify what are the most abundant class of transposable element and apply statistical tests to determine their relationships. I also identify retroposons, namely the Juan elements, which are present in unusually high copy numbers in the genomes. I then tested what is known about these elements based on previous publications. This chapter provides new molecular biology and bioinformatics data done on these elements.

Chapter 4: Characterisation of Culex pipiens 1, Pip1, an Active Low Copy Number Retroposon that has a Novel Start Codon

Here, I characterise the retroposon Pip1. Previous publications did not attempt to describe the element in detail. I use both molecular biology and bioinformatics approach to characterise the element, describing its phylogenetic position, copy number and important coding domains.

Chapter 5: Artificial Horizontal Transfer of Retroposons

This chapter describes the research done on using Pip1 and the Juan elements as germline transformation tools. Each element was inserted into a vector before being introduced into the germline of *D. melanogaster*. I then established a homozygous breeding line. Whole genomic sequencing data was obtained to determine insertion sites.

Chapter 6: Concluding Remarks

I summarise the main conclusions of the chapters. I also explore future directions of research in this area.

CHAPTER 2

GENERAL MATERIALS AND METHODS

2.1 Insect Material

2.1.1 Insect strains

The *Aedes aegypti* Liverpool strain was established and maintained at the Liverpool School of Tropical Medicine since 1936 and is the strain used as the reference (Nene *et al*. 2007).

Culex quinquefasciatus Muheza strain originates from Tanzania. It was collected from a sample of wild mosquitoes and established in 1986 (Khayrandish and Wood, 1993). The Johannesburg strain was established from Johannesburg, South Africa since 2001 and is the strain used as the reference genome (Arensburger *et al*. 2010).

The *Drosophila melanogaster* Canton-S strain is a well established wild type fruitfly stock. The yellow white strain was established from mutations causing white eye colouration and yellow body pigmentation (Santamaria, 1986).

2.1.2 Mosquito rearing

The mosquitoes were kept in a room with a constant temperature of 22°C and humidity of 70%. Eggs were hatched in bowls of water and larvae were fed with conventional fish food. Pupae were then transferred into cages where they emerge into adults. Adults were kept in 30cm x 30cm x 30cm cages and fed on 10% glucose solution. A blood meal was given to boost egg production. Eggs were collected in bowls lined with filter paper and either stored or allowed to hatch.

No animals were used for the blood feeding. The mosquitoes were fed on a a healthy volunteer (Dr. Colin Malcolm). His arms were placed on the side of the wire

mesh of the cage. Adult female mosquitoes were allowed to feed up to 2 minutes before the arm was removed. The volunteer had previous experience of feeding mosquitoes using this method and seems to have a low level of sensitivity. He did not suffer from any ill effects during the experiments.

2.1.3 Fruitfly rearing

Flies were kept at 70% humidity, 25°C room. Flies were kept in 7.5 x 2.25cm vials filled with ¼ cornmeal-treacle-agar medium. The flies were transferred into fresh vials about every two weeks.

2.1.4 Cornmeal-treacle-agar medium

100g agar, 150g sucrose, 350g D-glucose, 350g yeast, 150g maize meal, 10 tablespoons of soya flour, 300g treacle and 100g wheat germ was added to 10L of water. The solution was boiled and allowed to simmer for 10 minutes. The solution was cooled and 100ml of 10% Nipagin diluted in ethanol and 50ml of propionic acid was added to the solution. The solution was then dispensed into glass vials and plugged after solidifying.

2.1.4 Agar plate medium

25g sucrose and 10g agar was added and dissolved completely in 100ml of water. Next, 250µl phosphoric acid and 2.25ml propionic acid were added and filled up to 500ml of water. The solution was then dispensed into 3.5cm petri dishes.

2.1.5 Anaesthetizer

Flies were tipped into a carbon dioxide anaesthetizer and observed under a microscope. The flies were in the anaesthetizer for a maximum of 20 minutes before transferred back into food vials.

2.1.6 DNA extraction

This method was used to prepare both mosquito and fruitfly DNA. Single or multiple insects were homogenised in a 1.5ml eppendorf tube suspended in 100µl extraction buffer (1% SDS, 50mM TrisHCl at pH 8.0, 25mM NaCl and 25mM EDTA at pH 8.0). The samples were then incubated at 68°C for 15 minutes. Next, 100µl of 3M potassium-acetate at pH 7.2 was added, and the tubes were left on ice for 15 minutes. The samples were spun in a table-top centrifuge at 13000 rpm for 5 minutes. The supernatant was decanted into a fresh eppendorf tube and the pellet discarded. 600µl of ice cold 100% ethanol was added to the supernatant and left to precipitate for a minimum of 2 hours at -20°C. The tubes were then centrifuged for 10 minutes and the supernatant discarded. 100µl of ice cold 70% ethanol was added and the tubes spun for 5 minutes. The previous step was repeated. 100µl of ice cold 100% ethanol was added and the tubes centrifuged again for 5 minutes. The supernatant was discarded and the tubes left to dry to drain the supernatant. Finally, the pellet was resuspended in 20µl of sterilised distilled water. To remove RNA, 0.5µl of RNase H [New England Biolabs (UK) Ltd] was added and incubated at 37°C for 20 minutes and subsequently at 65°C for 20 minutes.

2.2 Gel electrophoresis

2.2.1 DNA Gels

DNA samples were mixed with 10x gel loading buffer [6x bromophenol blue 0.25% (w/v)] and loaded onto 0.8% (w/v) agarose gels (Bioline) made with 1x TAE buffer (40mM Tris-acetate, 1mM EDTA, pH 8.0). Gels were run at 95V for 2 hours using a horizontal gel apparatus (Perfect Blue Gel System Midi ExW, PEQLAB). The size markers used was Hyperladder I (Bioline). Gels were stained with 5µl ethidium bromide for 20 minutes and visualised under UV light and photographed.

2.2.2 Gel Extraction

The MinElute Gel Extraction kit (Qiagen) was used to extract DNA from agarose gels according to the manufacturer's instructions. Briefly, the DNA fragment was excised with a scalpel from the agarose gel. 3 volumes of buffer QG was added to 1 volume of gel. The gel slice was incubated at 50°C for 10 minutes or until it has completely dissolved. The tube was inverted every 3 minutes to aid the process. 1 gel volume of isopropanol was added to the sample. The sample was loaded into a MinElute column and centrifuged for 1 minute at 13 000rpm. Elute was discarded and 500µl of buffer QG was added and centrifuged for 1 minute at 13 000rpm. 750µl of buffer PE was added and centrifuged for 1 minute at 13 000rpm. The elute was discarded and the column was centrifuged again for 1 minute at 13 000rpm. The columns were placed in eppendorf tubes and 10µl prewarmed (50°C) sterile distilled water was added to the column. The column was left to stand for 1 minute before centrifuged for 1 minute at 13 000rpm. The elute was collected and 1µl was analysed on an agarose gel.

2.3 Polymerase Chain Reaction (PCR)

2.3.1. Primer Design

Primers were designed using Primer-BLAST (Rozen and Skaletsky, 2000)

2.3.2 Standard PCR

A concentration of 1x Standard *Taq* Reaction Buffer, 200µM of each dNTP, 0.2µM of each primer and 1.25 units of Taq polymerase was used. The steps used in the PCR was initial denaturation at 95°C for 30 seconds, 30 cycles of denaturation at 95°C for 15 seconds, annealing for 30 seconds and elongation at 68°C for 45 seconds per kb, and a final elongation step at 68°C for 5 minutes. The PCR machine used was peqSTAR 96 Universal cycler (PEQLAB).

2.3.3 Expand Long Template PCR

The Expand Long Template PCR system (Roche) was used to amplify long DNA fragments. A concentration of 1x Expand Long Template Buffer 1, 350µM of each dNTP, 0.3µM of each primer and 0.5 units of Expand Long Template Enzyme DNA polymerase was used. The steps used in the PCR was initial denaturation at 92°C for 2 minutes, 10 cycles of denaturation at 92°C for 10 seconds, annealing for 30 seconds and elongation at 68°C for 45 seconds per kb, followed by 20 cycles of denaturation at 92°C for 15 seconds, annealing for 30 seconds and elongation at 68°C for 45 seconds per kb, and a final elongation step at 68°C for 7 minutes. The PCR machine used was peqSTAR 96 Universal cycler (PEQLAB).

2.3.4 Phusion Polymerase and Site-Directed Mutagenesis

Phusion Polymerase was used to correct changes in the DNA sequence. A concentration of 1x Phusion HF Buffer, 200µM of dNTPs, 0.5µM of each primer and 1 unit of Phusion Polymerase was used. The steps used in the PCR was initial denaturation at 98°C for 30 secodns, 30 cycles of denaturation at 98°C for 10 seconds, annealing for 30 seconds and elongation at 72°C for 30seconds per kb, with a final elongation step at 72°C for 10 minutes. The PCR machine used was peqSTAR 96 Universal cycler (PEQLAB).

2.3.5 PCR purification

The MinElute PCR Purification kit (Qiagen) was used to purify PCR products according to the manufacturer's instructions. Briefly, 5 volumes of buffer PBI was added to 1 volume of PCR reaction and mixed. The sample was then applied to the MinElute columns provided and centrifuged for 1 minute at 13 000rpm. The elute was discarded and 750µl buffer PE was added and centrifuged for 1 minute at 13 000rpm. The elute was discarded and the column was centrifuged again for 1 minute at 13 000rpm. The columns were placed in eppendorf tubes and 10µl prewarmed (50°C) sterile distilled water was added to the column. The column was left to stand for 1

minute before centrifuged for 1 minute at 13 000rpm. The elute was collected and 1µl was analysed on an agarose gel.

2.4 Cloning of PCR products

2.4.2 Restriction enzyme digests

All restriction enzymes were obtained from NEB (New England Biolabs, UK). All reactions were optimised using their corresponding buffers. A unit of enzyme is used per reaction and left to incubate for a minimum of 2 hours at the enzyme optimum temperature.

2.4.1 Ligation

1x T4 DNA Ligase Buffer and 3 units of T4 DNA ligase (Promega) were used. A ratio of 1:2 to 1:3 of vector:insert was used in the ligation mix. The reactions are incubated for a minimum of 1 hour at room temperature or overnight at 4°C.

2.4.2 Transformation

2µl of the ligation product was transferred to an eppendorf tube. 50µl of JM109 High Efficiency Competent Cells (Promega, USA) was added and gently mixed by flicking the tube. The reaction was left in ice for 20 minutes, heat-shocked for 45 seconds at 42°C and left in ice for 2 minutes. 950µl of SOC medium was added and the reaction incubated for 1.5 hours at 37°C with shaking at 150rpm. 100µl of the mix was plated onto LB/ampicillin (10mg/ml) plates and incubated overnight for selection.

Colonies were screened using PCR. Colonies which gave a positive PCR reaction was then harvested for their plasmids and restriction enzyme checked. Double positive colonies were kept in LB Broth with 25% glycerol and stored in the -80°C freezer.

2.4.3 Plasmid preparation

Plasmids were harvested using Qiaprep Spin Miniprep Kit (Qiagen) according to the protocol supplied. *E. coli* cultures were grown in 5ml LB medium containing 20μg/ml ampicillin overnight for 16 hours at 37°C prior to harvesting. The cells were centrifuged at 4000g for 10 minutes at 4°C. The supernatant was discarded and the pellet was resuspended in 250μl buffer P1 and transferred to an eppendorf tube. 250μl buffer P2 was added and inverted 6 times. 350μl buffer N3 was added and inverted 6 times. The sample was centrifuged at 13 000 rpm for 10 minutes. The supernatant was applied to a QIAprep spin column and centrifuged at 13 000 rpm for 1 minute. The elute was discarded and 0.5ml buffer PB was added to the column and centrifuged at 13 000 rpm for 1 minute. The elute was discarded and 0.75ml buffer PE was added to the column and centrifuged at 13 000 rpm for 1 minute. Elute was discarded and the column was centrifuged at 13 000 rpm for 1 minute. The columns were placed in eppendorf tubes and 50μl prewarmed (50°C) sterile distilled water was added to the column. The column was left to stand for 1 minute before centrifuged for 1 minute at 13 000rpm. The elute was collected and 1μl was analysed on an agarose gel.

2.4.4 DNA sequencing

DNA sequencing was carried out by Eurofins MWG Operon (London, UK). Whole genomic sequencing using MiSeq was performed by the Genome Centre (Charterhouse Square, Queen Mary).

2.5 Bioinformatics

2.5.1 Databases

The database Repbase (Jurka et al, 2005), GenBank (NCBI) and VectorBase (Megy *et al*, 2009) was used.

2.5.2 Bioinformatics programmes

BLAST (Zhang *et al*, 2000) was used to obtain homologus sequences.

DNA sequence alignments and phylogenetic trees were drawn using CLC DNA Workbench Version 6.0.2 (CLC Bio, Denmark).

Repeat Masker was run on http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker to estimate the number of retroposon copies in the genome (Smit *et al*, unpublished).

Statistical tests were carried out using the statistics package R (The R Core Team, 2015)

CHAPTER 3

RETROPOSONS AND THE MOSQUITO GENOME

3.1 Introduction

Transposable elements (TEs) are ubiquitous: Since their discovery by Barbara McClintock (1956) in maize, TEs have been found in most organisms, from prokaryotes to eukaryotes. In the era of genome sequencing, the discovery of new TEs has been facilitated by improvements in sequencing technology as well as better bioinformatics programmes to detect TEs (Durand *et al*, 2006; Janicki *et al*, 2011).

Our perception of TEs has also changed over time. Barbara McClintock called the mobile elements she found 'controlling elements' – due to the fact that kernel colours were different depending on the activity of genes at the dissociator and activator loci. Then, TEs gained the tag of being selfish, parasitic and junk; a perspective which was popularised by Richard Dawkins in his book The Selfish Gene (1976). TEs were viewed as not conferring any benefit to the host genome at all: rather, as excess baggage in the genome, detrimental to the host.

For a time, TEs were characterised as having a negative impact on the host genome because TE activity is a major source of genome mutation. Direct insertion into exons will produce corrupt translation of the protein product. P elements in *Drosophila melanogaster* are famously known for producing mutant phenotype flies in addition to causing hybrid dysgenesis (Rubin and Spradling, 1982). Transposition also disrupts gene regulation because insertions upstream and downstream alter gene expression (Morgan *et al*, 1999). A high copy number of a single element also promotes non-homologous recombination (Kidwell and Lisch, 2001; Oliver and Greene, 2009). Two elements residing in the same or different chromosomes could pair up and

recombine, causing genetic information to be gained or lost in the chromosomes. This effect is analogous to the problem that TEs pose for bioinformatics whereby the match between two TEs in different positions can lead to the incorrect assembly of a shortened contig – skipping the sequence between the two TEs (Waterhouse *et al*, 2008). TE activity can also produce pseudogenes (Kidwell and Lisch, 2001). There is a category of pseudogenes that resemble mRNA transcripts of genes: they do not contain any introns or regulatory elements and have a poly-adenosine tail. These properties would be explained if the pseudogene arose by reverse transcription of messenger RNA, which could be carried out the by reverse transcriptase from TEs. Therefore, having high amounts of TEs could be very damaging to the genome.

However, in addition to these negative (or at most neutral) effects, there is evidence that TEs can play an important role in the host genome. Very rarely, a TE insertion which produces an altered gene product might benefit the host (Darboux *et al*, 2007). The telomeres of *D. melanogaster* are made up of Het-A, TAHRE and TART retroposons and without the activity of these elements, *D. melanogaster* would not be able to produce telomeres (Mason *et al*, 2007; Shpiz *et al*, 2007). TEs might also play an important role in speciation. TE activity in bats coincides with the explosion of bat speciation (Ray *et al*, 2008). Promoter regions of TEs could also be recruited to regulate host genes (Gonzalez and Petrov, 2009). With more research and as new evidence comes to light, our perception of TEs will probably change again.

Since the initial sequencing of the human genome (IHGSM, 2001) the estimate of the TE composition in the human genome has varied from 50% to 70% (de Koning *et al*, 2011). This leads to the question of why does the human genome have such a high TE content, since it is potentially damaging, and how did it happen in the first place? One clue could be the fact that most of the TE content is due to a single type of element: LINE-1. The human LINE-1 family arose ~4MYA and has been the dominant TE in the human genome (Boissinot *et al*, 2000; Hancks and Kazazian, 2012). The high contribution to the human genome could therefore be related to a recent spread

through the genome. This explanation is supported by evidence that LINE-1 is still active and insertions are polymorphic between different populations (Beck *et al*, 2010). In addition, another mobile genetic element, Alu, exploits LINE-1's transposition machinery to generate copies of itself. Therefore, the abundance of TEs in the human genome could be predominantly due to the activity of LINE-1.

Various groups have agreed upon the phylogeny relationship between mosquitoes and the phylogeny relationship is represented in Figure 3.1. (Rao and Rai, 1987; Miller *et al*, 1997; Rai, 2010; Vicoso and Bachtrog, 2015). A chaoborid midge was the closest relative to the mosquitoes, and the mosquitoes formed a monophyletic group. Within the mosquitoes, *Anopheles* is places as the outgroup to *Culex*, *Aedes* and *Toxorhynchites. Toxorhynchites* is placed as the outgroup to *Culex* and *Aedes*.

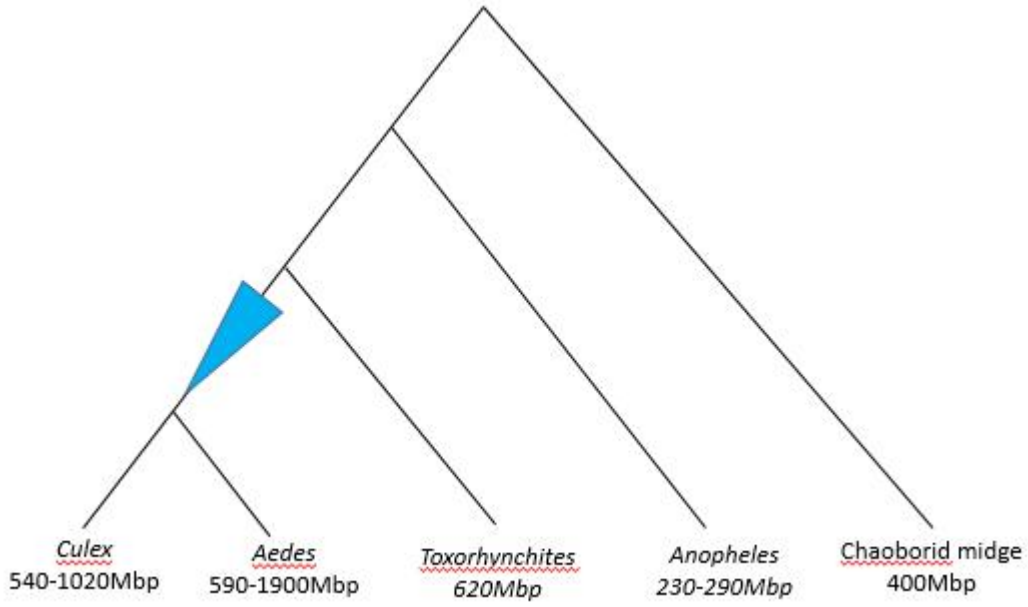

Figure 3.1 A cladogram including the key insect taxa. Genome size ranges are indicated under each taxon name (values from Rai 2010). The blue triangle indicates where the increase in insect genome size is proposed. N.b. branch lengths are not proportional to time.

In light of the established phylogenetic relationships, the most parsimonious explanation is one episode of increase in genome size (shown in Fig 3.1). The ancestral genome under this scenario would have been small, with a small proportion of TEs. The increase in genome size occurred in the *Culex-Aedes* ancestor. This could have occurred along with an increase in abundance of TEs. Rai (2010) has argued that the alternative scenario of a large ancestral mosquito genome is unlikely, as this would require many reductions in genome size in independent evolutionary branches.

The Class I retroposons make up the highest autonomous TE content in each mosquito genome. 'Retroposons' is a term used to describe Class I TEs that mobilize through an RNA intermediate (Eickbush and Malik, 2002). In particular, a highly repetitive retroposon is present in the *Ae aegypti* and *Cx. quinquefasciatus* genome. JuanA makes up to 3% of the *Ae. aegypti* genome (Biedler and Tu, 2007; Mouches *et al*, 1992). It has recently been active in evolutionary time. A full length element is 4709bp long and has an internal promoter, two open reading frames (ORFs), and a poly-adenosine tail. Insertions of the element produce target site duplications. The first reading frame encodes a cysteine rich domain while ORF2 encodes an endonuclease as well as reverse transcriptase which are vital for activity. A phlyogenetic analysis by Crainey *et al*. (2005) suggests that this element has survived via vertical transfer and not horizontal transfer. In addition, it has a high ratio of dS/dN (10.7+/- 2.9) (Biedler and Tu, 2007); an observation which supports the idea that it is under evolutionary pressure to remain functionally active.

JuanC is the closest relative of JuanA but is present in *Cx. pipiens* (Agarwal *et al*, 1993). It is 4.48kb long and also contains an internal promoter, two ORFs and a poly-adenosine tail. Target site duplications indicate that an insertion via reverse transcription has occurred in the genome. It has been estimated that the haploid *Cx. pipiens* genome contains 2500 full length elements. Both Juan elements belong to the Jockey clade which is only present in insects. The amino acid sequences of JuanC ORF1

and ORF2 share 39.5 and 66.9% homology respectively with JuanA (Agarwal *et al*, 1993).

The contribution that TEs make towards genome sizes in mosquitoes is examined here as a possible paradigm, because more TEs effectively reduces gene density thereby presumably reducing the risk of negative impact due to transposition and allowing even more expansion. The analysis is extended to focus on Juan, as the emergence of the large Juan families against a background of many diverse TE families almost certainly reflects an earlier stage to the situation in humans, so helping to explain how or why the human genome came to be so dominated by LINE-1. The objective was to identify features that might have contributed to their success in out-competing other TEs.

One not so obvious question was what is the sequence of a functional element capable of transposition? An initial bioinformatics analysis indicates that the published JuanA and JuanC sequences are not completely representative of typical elements. A combination of *in silico* and experimental procedures were used to obtain more robust sequence data, which provided the template against which successfully cloning of full length and potentially active elements in the genome was measured. The final objective of this section of the work was to obtain full length and potentially active sequences of JuanA and JuanC in clones.

3.2 Materials and Methods

3.2.1 Insect genome sizes and TE content

Information concerning the number of TEs and the genome sizes were obtained from various published genome sequencing projects. The genome size, percentage of TE as well as number of protein coding genes were obtained from the respective genome sequencing paper (referenced in Table 3.2). Briefly, the TE content and protein

coding genes were identified using bioinformatics programmes that search for similarities to known sequences as stated in their respective genome sequencing paper. The data was tabulated and graphs were drawn in Microsoft Excel. A graph of percentage of TE composition in genome was plotted against genome size (Figure 3.2).

A correlation test was carried out on the data. In order to allow for the dependency in the data due to shared common ancestry, each insect was paired with its closest evolutionary relative and the difference between genome size and difference in amount of TE sequence were calculated, in mega base pairs; the correlation across pairs and statistical signficance were calculated using the cor.test() function of R (The R Core Team, 2015).

The same design was also used test the relationship between genome content and number of different classes of TE in the mosquitoes using for the two *Anopheles* and difference, *Culex* and *Aedes*.

3.2.2 Bioinformatics

The genome sequences were mined using BLAST (Zhang *et al*, 2000) on the NCBI website, using the reference genomic sequences as database. All parameters were set to default except that the target sequence parameter was set to 1000 because of the high Juan copy number. The sequence of JuanA can be found at accession number M95171 while JuanC can be found at M91082 on the NCBI website. The JuanA BAC clones used in this study were BAC ND41B18 (Acc. number EF173373.1), BAC 105H24 (Acc. number EF173366.1) and XX-10B1 (Acc. number AC150259.4).

Alignments were carried out using CLC DNA Workbench Version 6.0.2 (CLC Bio, Denmark). Repeat Masker was run on http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker (Smit *et al*, unpublished).

3.2.3 Mosquito DNA extraction

Single mosquitoes (*Ae. egypti* Liverpool strain or *Cx. quinquefasciatus* Johannesburg strain) were homogenised in a 1.5ml eppendorf tube suspended in 100µl extraction buffer (1% SDS, 50mM TrisHCl at pH 8.0, 25mM NaCl and 25mM EDTA at pH 8.0). The samples were then incubated at 68°C for 15 minutes. Next, 100µl of 3M potassium-acetate at pH 7.2 was added, and the tubes were left on ice for 15 minutes. The samples were spun in a table-top centrifuge at 13000 rpm for 5 minutes. The supernatant was decanted into a fresh eppendorf tube and the pellet discarded. 600µl of ice cold 100% ethanol was added to the supernatant and left to precipitate for a minimum of 2 hours at -20°C. The tubes were then centrifuged for 10 minutes and the supernatant discarded. 100µl of ice cold 70% ethanol was added and the tubes spun for 5 minutes. The previous step was repeated. 100µl of ice cold 100% ethanol was added and the tubes centrifuged again for 5 minutes. The supernatant was discarded and the tubes left to dry to drain the supernatant. Finally, the pellet was resuspended in 20µl of sterilised distilled water. To remove RNA, 0.5µl of RNase H [New England Biolabs (UK) Ltd] was added and incubated at 37°C for 20 minutes and subsequently at 65°C for 20 minutes.

Ae. aegypti Liverpool strain and BAC clone ND41B18 was obtained from the genome sequencing project (Dr. Ranson, Liverpool School of Hygiene and Tropical Medicine, UK). The *Cx. quinquefasciatus* Johannesburg strain was obtained from Niki Pool from University of California, USA.

3.2.4 PCR

Various primers were designed using Primer-BLAST to amplify both Juan sequences (Table 3.1) (Rozen and Skaletsky, 2000). Expand™ Long Template PCR system (Roche, Germany) was the PCR system used to obtain the full length element. The forward primers JA34F, JA36F, JA40F, JA42F, JA43F and JA45F are internal primers close to the 5′ end, while the reverse primer JA4545R is an internal primer close to the 3′ end. These primers would produce 4.5kb products without the 5′ and 3′

ends of JuanA. The forward primer JAfl5F and JABAC92828F flanks the 5′ end of JuanA while JuanAfl3AR and JABAC9787R flanks the 3′ end of JuanA- this would produce the full length 4.7kb JuanA element with additional flanking genomic sequences.

The forward primer JCfl5F flanks the 5′ end of JuanC while JuanCfl3R flanks the 3′ end of JuanC- this would produce the full length 4.6kb JuanC element with additional flanking genomic sequences.

Table 3.1. PCR primers used to in this study. JuanA primers are named with the prefix JA while JuanC are named with JC.

Forward primer	**DNA sequence**	**Reverse primer**	**DNA sequence**	**Annealing temperature**
JA34F	ACGAATTCTCTCTG CTCTTG	JA4545R	GTGAGTTGATTTCA CCTGCT	50
JA36F	GAATTCTCTCTGCTC TTGGA	JA4545R	GTGAGTTGATTTCA CCTGCT	50
JA40F	CTCTGCTCTTGGAA GTT	JA4545R	GTGAGTTGATTTCA CCTGCT	50
JA42F	TCTCTGCTCTTGGAA GTTTT	JA4545R	GTGAGTTGATTTCA CCTGCT	50
JA43F	CTCTGCTCTTGGAA GTTTTC	JA4545R	GTGAGTTGATTTCA CCTGCT	50
JA45F	CTGCTCTTGGAAGTT TTCTT	JA4545R	GTGAGTTGATTTCA CCTGCT	50
JAfl5f	ACGCTTACGCCTTG AAAATG	JAfl3Ar	CGAACGATGAACA AAAATCG	52
JAF	CCTTTCGAAGGTCA CGTCTT	JA1880R	CCATTCAGAGAAC GAGCATT	53
JABAC 92828F	GGAAGTCCCAAGGA GGTTTT	JABAC 97987R	CGATATTGAAGGG ACCATCG	54
JC 2F	TGACCTCAAAACGG ACAGTCT	JC 4625R	CTCAGCCATAACA TGGTGGTT	57
JCfl5F	CTCGTCGACAAATG CGTCAAAACAAAG	JuanCfl3R	TATGGAAAGAGAG AGTGCAAAGC	59

Expand™ Long Template PCR system (Roche, Germany) was the PCR system used to obtain the full length element. Each PCR reaction contained 10µM of primers; 500µM of dNTPs; 5µl of 10x Expand Long Template Buffer 1 (17.5mM $MgCl_2$); 0.5 units of Expand™ Long Template Enzyme mix. The PCR programme was (1) Heat PCR machine to 120°C. (2) 93°C for 2 minutes. (3) 10 cycles of 93°C for 10 seconds, annealing temperature (see Table 3.1) for 30 seconds, and 68°C for 4 minutes (4) 25 cycles of 93°C for 15 seconds, 50°C for 30 seconds, and 68°C for 4 minutes + 20 second for each successive cycle. (5) Final elongation at 68°C for 7 minutes. PCR products were analysed on an agarose gel.

3.2.4 Sequencing of PCR products

PCR products were purified using QIAquick® PCR Purification Kit (QIAGEN, UK) following the protocol provided by the supplier using a microcentrifuge. Promega pGEM® T Easy Vector system was used to transform J109 *E. coli* competent cells (Promega, USA) following the protocol supplied by the manufacturer. Cells were grown on plates containing LB broth, ampicillin, IPTG and X-Gal. Glycerol stocks were kept in the freezer at -80°C. Plasmid were harvested using QIAprep Spin Miniprep Kit (QIAGEN, West Sussex, UK) and a microcentrifuge, according to the protocol supplied by the manufacturer. DNA sequencing was done by Eurofins MWG Operon (London, UK).

3.3 Results

3.3.1 Insect genome sizes and TE content

The genome sizes and TE content of the mosquitoes were compared with different sequenced insects (Table 3.2). The genome size (mega base pairs, Mbp) of each insect species with the amount of TE composition is presented in Figure 3.2. The correlation test demonstrated that genome size and TE content are highly correlated ($t = 18.5729$, $df = 4$, p-value = 4.9×10^{-5}).

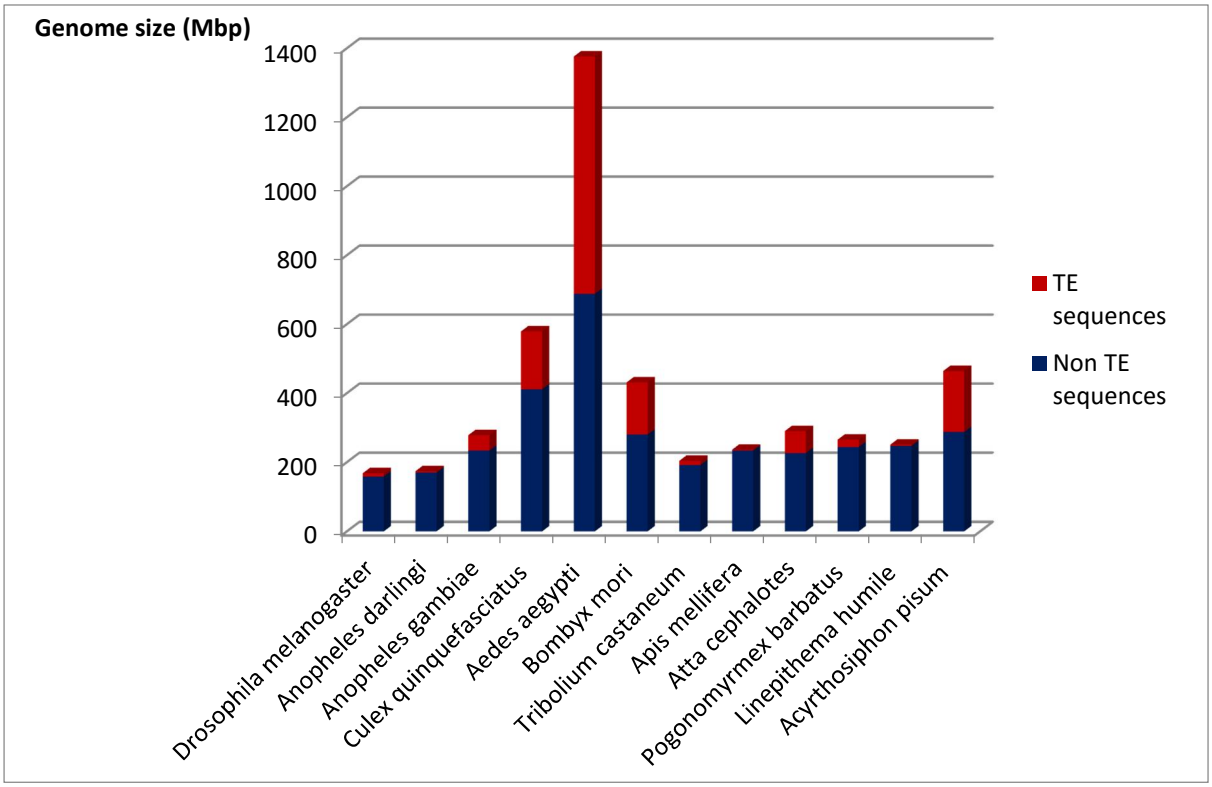

Figure 3.2. The genome sizes of different sequenced insect genomes and the proportion of TEs in these genomes. Non-TE sequences are in blue while TE sequences are in red. Refer to table for references

Table 3.2. The genome sizes, composition of TEs and number of protein-coding genes in the genomes of insects. All data was collected from the respective genome sequencing projects.

Species	Genome size (Mega base pairs)	Percentage of transposable elements (%)	TE content (Mbp)	Number of protein-coding genes
Drosophila melanogaster, Fruitfly (Adams *et al*, 2000; Lee and Langley, 2010)	168	5.5	9.24	13 601
Anopheles darlingi, Neotropical malaria vector (Marinoti *et al*, 2013)	174	2.3	4.002	10 457
Anopheles gambiae, Malarial mosquito (Holt *et al*, 2002)	278	16	44.48	12 457
Culex quinquefasciatus, Southern house mosquito (Arensburger *et al*, 2010)	579	29	167.91	18 883
Aedes aegypti, Dengue fever mosquito (Nene *et al*, 2007)	1376	50	688	15 419
Bombyx mori, Silkmoth (International Silkworm Genome Consortium, 2008; Osanai-Futahashi *et al*, 2008)	431	35	150.85	16 329
Tribolium castaneum, Flour beetle (Tribolium Genome Sequencing Consortium, 2008)	204	6	12.24	16 400
Apis mellifera, Honeybee (Honeybee Genome Sequencing Consortium, 2006)	236	1	2.36	17 000
Atta cephalotes, Leaf cutter ant (Suen *et al*, 2011)	290	21.9	63.51	18 093
Pogonomyrmex barbatus, Red harvester ant (Smith *et al*, 2011)	265	7.93	21.0145	17 177
Linepithema humile, Argentine ant (Smith *et al*, 2011)	250.8	1.4	3.5112	16 123
Acyrthosiphon pisum, Aphid (The International Aphid Genomics Consortium, 2010)	464	38	176.32	32 800

An analysis of the different classes of TEs in the mosquitoes is presented in Figure 3.3. Autonomous and non-autonomous Class I and Class II TEs are present in the entire mosquito group, but the most abundant TEs are retroposons. The correlation test performed was found to be statistically significant (t = 5.6114, df = 5, p-value = 0.002486). The increase in genome size is accompanied by an increase in amount of TE sequences as well, from *An. darlingi* to *An. gambiae* to *Cx. quinquefasciatus* to *Ae. aegypti*. The Class I retroposons are the most abundant autonomous TEs in the genomes.

However, the correlation between the genome size and the number of genes found that the correlation between them is not significant (p-value = 0.7361).

The sequencing of different insect genomes has allowed a better analysis of change in genome size and TE content. The correlation test for genome size and amount of TE found that there is a statistically significant relationship (p-value = 4.9×10^{-5}). When genome size increased, there was an increase in TE content as well. However, this does not imply a causative relationship. Each could increase independently of each other. Genome size could have increased via other processes such as gene duplication. TE content could increase by uncontrolled mobilisation. Among the mosquitoes, the genome size and the number of different TE classes is also strongly correlated (p-value = 0.002486). Bigger genomes have more classes of TE in their genomes.

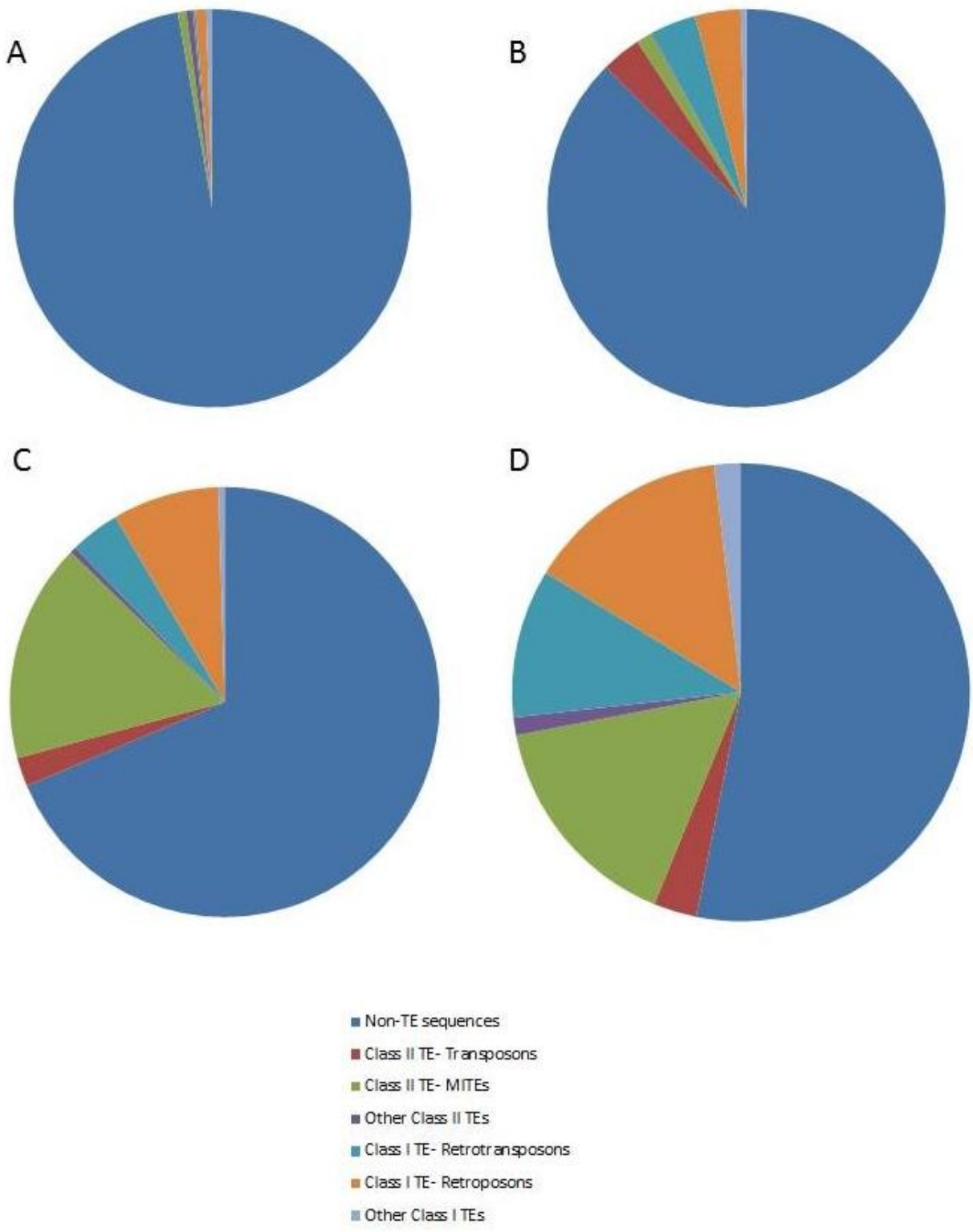

Figure 3.3 Comparison of the genomic composition of the four sequenced mosquito genomes. It includes both autonomous and non-autonomous TEs A) *An. darlingi*; B) *An. gambiae*; C) *Cx. quinquefasciatus*; D) *Ae. aegypti*.

3.3.2 Reanalysis of JuanA DNA Sequence

The JuanA DNA sequence (accession number M95171) (Mouches et al. 1992), was queried on BLASTN against the Reference genome (Refseq) and against the Whole genome Shotgun (WGS) database for *Ae. aegypti* (taxon 7159). No identical copies were obtained, however it is necessary to display at least 1000 alignments to obtain sequences showing less than 99% identity and 99% query coverage and over 5000 to get an E value greater than zero. Even with BLASTN searches of the Nucleotide collection (nt) database no matches were obtained to sequences from any insect other than *Ae. aegypti*.

To obtain a reference sequence that would be representative of functional elements, 125 full length sequences from the *Ae. aegypti* WGS database were aligned. All sequences with insertions and most with deletions greater than one were then removed from the alignment. In making these cuts two mononucleotide tracts close to each other starting at positions 4517 and 4559 (see Figure 3.6) were ignored, as each varied from 6 to 12 nucleotides. This left 103 sequences, which were at least 99% identical to one another. Almost all of the variation involved random point mutations, with no more than one substitution in any column of the alignment. Where multiple substitutions were observed within a column it was normally the same substitution and only in a few sequences, but at 17 positions an alternative base was present in 10 to 29 sequences. All of the 1% variation was in non-coding regions. Only one position (4516) immediately prior to the poly-T tract mentioned above was more variable with a G or a T instead of an A in 37 and 14 sequences respectively. The alternative bases are diagnostic of derivation from a common progenitor, but a phylogenetic tree based on genetic distance produces sixteen monophyletic groups (results not shown), from which it would be difficult to identify one that was most recent.

A BLASTN search of the WGS database search with the majority consensus sequence did produce matches with 100% identity, but not better than 99% query

coverage. It is clear that while the *Ae. aegypti* genome contains an abundance of full length and near identical Juan elements, the majority are sufficiently old to have acquired unique point mutations, so it is difficult to predict the sequence of a fully functional element with complete certainty, nevertheless a consensus sequence was taken as a better approximation and used for the basis of further work. This analysis was not taken further as an extensive *in silico* analysis of the JuanA has already been conducted (Biedler and Tu, 2007).

The next objective was to obtain a full length copy of the element for cloning and subsequent analysis. Initial efforts to amplify the element by PCR using genomic DNA did not succeed despite, or perhaps because of, the high copy number in the genome. A series of PCR reactions in which the primer for the 5′ terminal varied produced multiple products except for 45f, which gave the expected result (Figure 3.4). Unfortunately this result proved difficult to reproduce and efforts to amplify from the initial product with the same primers produced a ladder of smaller products. The same *Ae. aegypti* strain (Liverpool) used in the genome sequencing project was used for the DNA template, so it was surprising that five of the six 5′ primers were apparently mispriming.

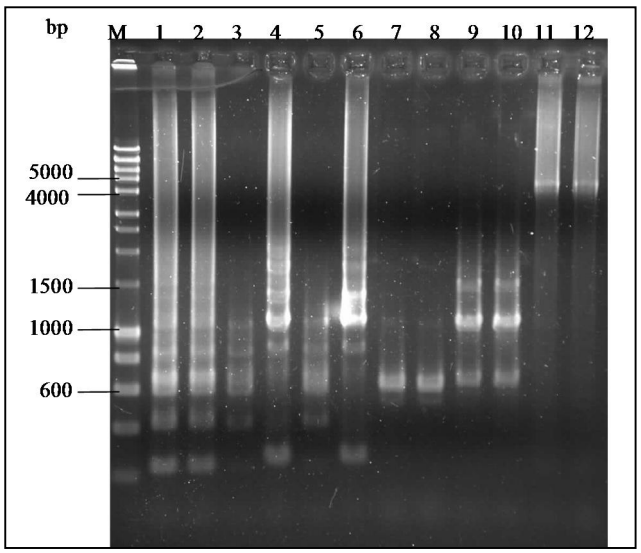

Figure 3.4 Gel electrophoresis result of PCR run on different sets of primers to amplify JuanA using whole genomic *Aedes aegypti* DNA. 1,2 -Primer 34f; 3,5- Primer 36f; 4,6- Primer 40f; 7,8- Primer 42f; 9,10- Primer 43f; 11,12- Primer 45f. The reverse primer used was 4545r. The expected 4.5kb product band was only present in in lanes 11 and 12. M is the Bioline HyperLadder 1 (Bioline, UK). The 1.5% agarose gel was ran at 90V for 4 hours.

With such an abundance of Juan elements, the use of genomic DNA presents two problems. The primers will anneal to truncated elements and if one terminal is present in excess it will deplete that primer relative to the other. Synthesis of truncated single strand Juan fragments from may anneal to single strand full length elements, thus interfering with subsequent rounds of amplification. Therefore a switch was made to using BAC clones from the genome sequencing project. A BAC clone will still contain multiple Juan elements, but it offered the opportunity of using primers based on the flanking sequence. At the time, only three BAC clones were available that contained full length JuanA elements and all three contained mutations on the second open reading frame (ORF) that gave a premature stop codon.

DNA from BAC ND41B18 was used together with primers flanking the JuanA element: BAC 92828F and BAC 97987R. A 5.5kb product was obtained (Fig 3.5), which was subsequently cloned and sequenced. The DNA sequence of this JuanA is presented in Figure 3.6.

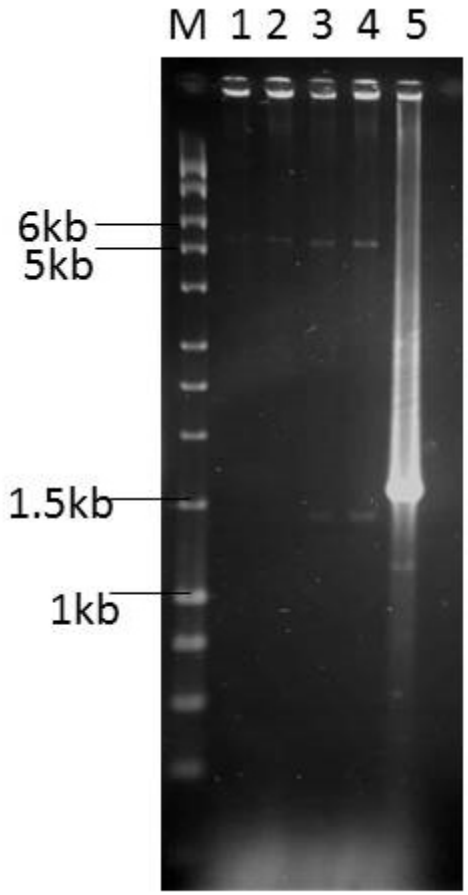

Figure 3.5 Gel electrophoresis result of PCR to amplify JuanA using flanking primers. The 5.5kb band was present in all lanes. Lane 5 is a positive control using primers internal to JuanA. M is the Bioline HyperLadder 1 (Bioline, UK). The 1.5% agarose gel was ran at 90V for 4 hours.

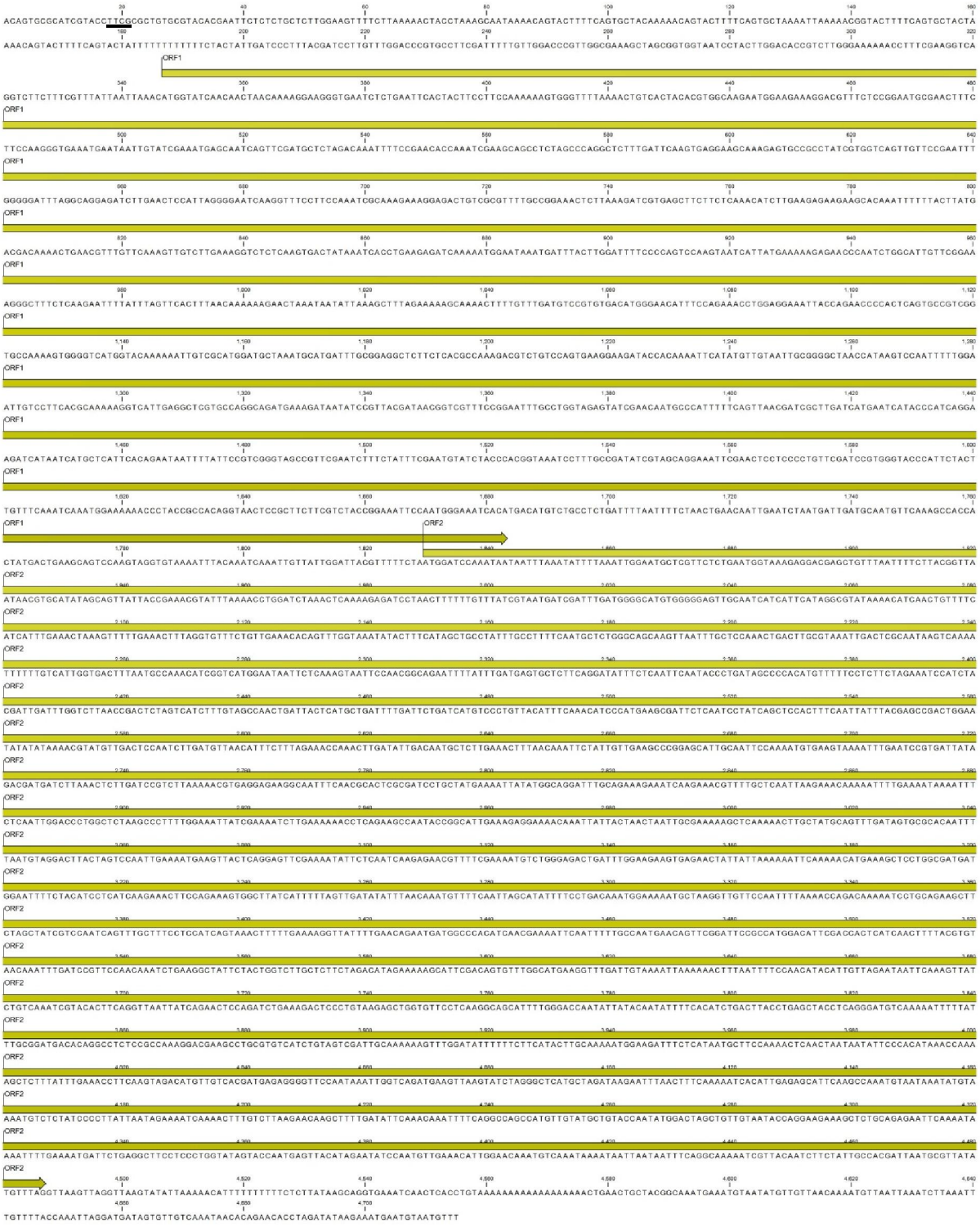

Figure 3.6 The DNA sequence of JuanA obtained via PCR cloning and sequencing from BAC clone ND41B18. The DNA sequence is shown in upper case letters The TTCG promoter sequence characteristic of Jockey elements is underlined. The ORF1 and ORF2 is shown in the diagram

The ORF2 in the published JuanA sequence (Mouches *et al*, 1992) does not have a methionine initiation of translation start codon. Instead, translation of ORF2 was hypothesized to occur via suppression of termination and template switching. However, the consensus based on 103 full length sequences described above and the sequences of the BAC clones possess an ATG start codon to initiate translation of ORF2. This difference is due to single nucleotide missing at position in M95171 (Figure 3.7). To further validate this, a BLASTN search using 200bp from the BAC clone sequence was conducted and gave 100% identity hits.

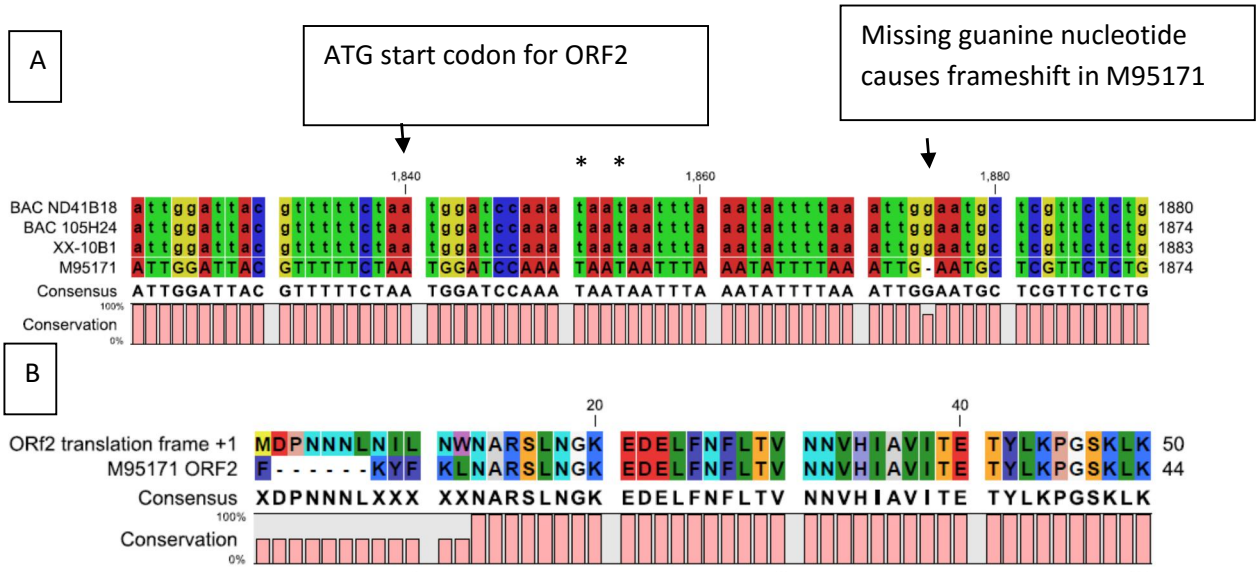

Figure 3.7 Differences in cloned JuanA with M95171. A) DNA alignment of the region at the 3′ end of ORF1 and 5′ end of ORF2. The two stop codons present directly after ORF1 is indicated by *. The ATG start codon for ORF2 as well as the missing nucleotide is also indicated in the diagram. B) Alignment of the amino acids of ORF2 at the 5′ end. The amino acid sequence is the same after the initial 12 amino acids.

3.3.3 Bioinformatics of JuanC DNA Sequence

The DNA sequence of JuanC is available from NCBI with the accession number M91082 (Agarwal *et al*, 1993). However, it is deposited together with the flanking DNA. Thus, the actual JuanC element starts 120 nucleotides downstream and ends 73 nucleotides earlier, with a total length of 4469 nucleotides. The DNA sequence and important features are highlighted and presented in Figure 3.8.

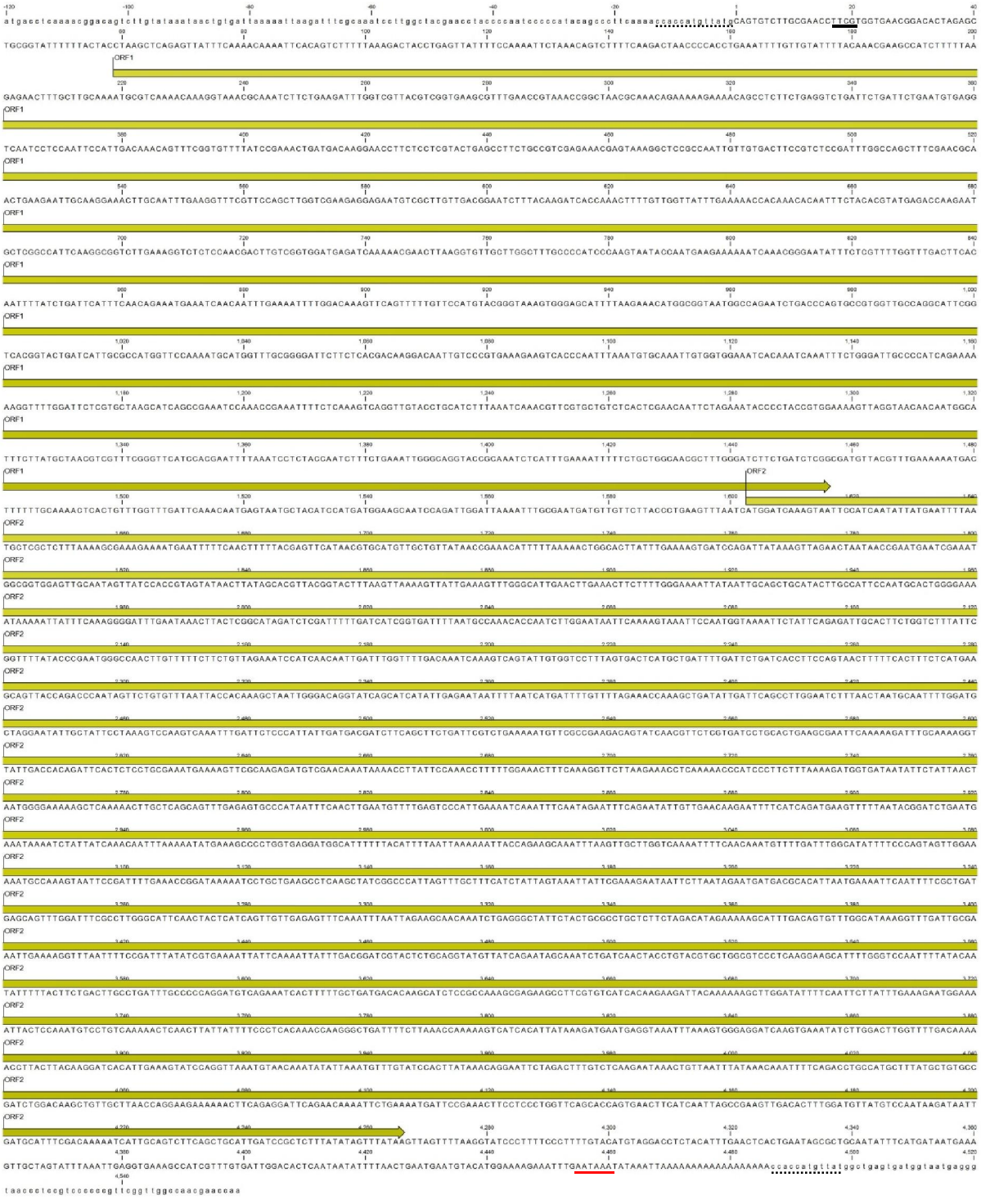

Figure 3.8 The DNA sequence of JuanC obtained via PCR cloning and sequencing from *Cx. quinquefasciatus* Johannesburg strain. The DNA sequence is shown in upper case letters. The TTCG promoter sequence characteristic of Jockey elements is underlined. Potential start codons are shown with arrows. ORF1 is highlighted with a yellow box while ORF2 is highlighted with a green box. The red line shows the polyadenylation signal, AATAA

It was estimated by Agarwal *et al* (1993) that as many as 2500 copies of JuanC are present in the haploid *Cx. pipiens* genome. However, using RepeatMasker, and sequences from the *Cx. quinquefasciatus* genome, there were only 1713 copies identified. These copies make up 0.6% of the *Cx. quinquefasciatus* genome.

3.3.4 Molecular biology of JuanC

A full length JuanC element was obtained via PCR. DNA from *C. quinquefasciatus* Johannesburg strain was used together with primers flanking the JuanC element. A 4.5kb product was obtained (Fig 3.9), which was subsequently cloned and sequenced.

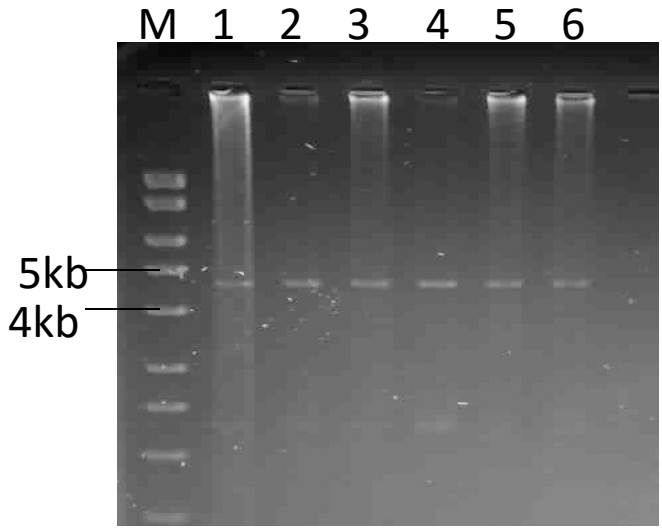

Figure 3.9 Gel electrophoresis result of PCR using primers near the ends of JuanC on the *Culex* mosquitoes. The band was successfully amplified in all of the PCR reactions. Expand Long Template PCR system was used. M is the Bioline HyperLadder 1 (Bioline, UK). The 1.0% agarose gel was ran at 90V for 4 hours.

3.3.6 Presence of triple CCHC motif at ORF1

The ORF1 of both elements contains 3 cysteine rich regions. A unique triple repeat of a cysteine rich region with a consensus of $CX_2CX_4H\ X_4C$-5aa- $CX_2CX_4H\ X_4C$- 9/10aa-$CX_2CX_3H\ X_6C$ is present. This consensus sequence corresponds to the Jockey zinc-knuckle domain. Previous publication only highlighted the first two cysteine rich

domain. The first two Cys rich domain is separated by 5aa while the last one is separated by 10 and 9aa in JuanA and JuanC respectively. The region from both elements, as well as Jockey and I factor elements from *Drosophila melanogaster*, was compared and a consensus sequence was obtained (Figure 3.10).

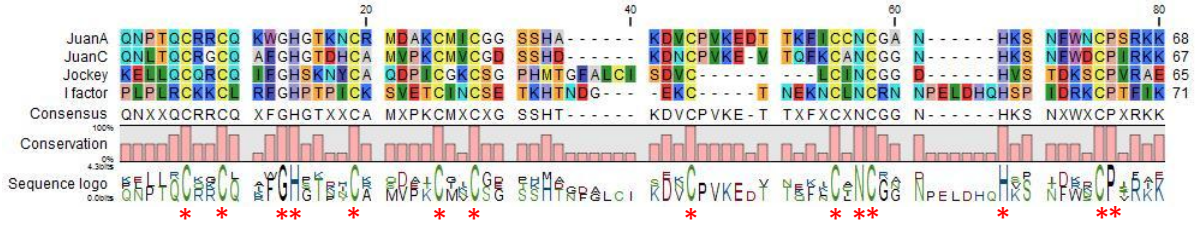

Figure 3.10 Alignment of the cysteine rich regions of the Juan elements, *D. melanogaster* Jockey element and *D. melanogaster* I factor. These elements possess a unique triple repeat of the CCHC motif, which is characteristic of zinc fingers. Additional 5 amino acids upstream and downstream are included in the alignment. Other amino acids appear to be conserved as well. Highly conserved regions are marked with a *.

The result was further validated by running the whole ORF1 on a protein prediction programme, Phyre[2] (Kelley and Sternberg, 2009). The zinc-finger region was modelled with >90% confidence.

3.4 Discussion

Most of the sequenced insect genomes have genome sizes less than 400Mbp and TE content less than 20%. However, the Dipterans, tend to have larger genome sizes. The correlation test demonstrated that genome size and TE content are highly correlated (p-value = 4.9×10^{-5}).

Charlesworth and Langley (1989) model the population genetics of TE copy number, assuming that the number of copies in the genome, n, is at an equilibrium between transposition, which increases the number of elements, and the deleterious effects of transposition, which select against the genomes with the largest copy number. They present an equation for the change in mean copy number in a population as:

$$\Delta\bar{n} \approx \bar{n}(\bar{n} - \bar{x})\frac{\partial \ln \bar{w}}{\partial \bar{n}} + \bar{n}\,(u_{\bar{n}} - v)$$

where $\bar{n}$ is the mean number of copies of TE per individual in a population, $\bar{x}$ is the mean frequency of elements in an occupable site, $\bar{w}$ is the mean of fitness of the host carrying n members of a given TE family (the differential term is the rate of change in mean fitness with the population-mean number of TE copies), $u_{\bar{n}}$ is the probability of transposition per generation and v is the probability of excision per generation. The first term on the left hand side of the equation describes the loss due to transposition while the second term describes the net gain in TE number.

Charlesworth and Langley suggest that the fitness, w, as a function of copy number $\bar{n}$ can be modelled by an exponential function: $w_n = \exp(-tn^2/2)$, where t is the slope of the relation between the logarithm of fitness and copy number at a copy number of 1). By inserting this into the above equation and solve for Δn=0, the equilibrium copy number is given as $\bar{n} = (u_{\bar{n}} - v)/t$.

If we wish to compare large and small genomes, it is reasonable to assume that the marginal effect of a new active TE would be lower in a large genome, because a smaller number of transposition effects hit vulnerable targets in the larger genome. Hence if the rate of transposition and excision remain the same, the copy number will be larger. In a bigger genome, there will be more potential sites that it could insert into. This larger number of active elements implies a lower fitness of larger genomes, since the mean fitness of the population at equilibrium, relative to the fitness of an element-free individual, is equal to $\exp(-n(u_{\bar{n}} - v))$ regardless of the form of the selection function (Charlesworth, 1985, cited in Charlesworth and Langley, 1989).

It was estimated that there are 378 functional JuanA copies (Biedler and Tu, 2007) while JuanC is estimated at 2500 copies (Agarwal *et al*, 1993). Using the same rate of transposition as Charlesworth and Langley (of the order of 10^{-4} per copy per generation), the fitness would be 0.96 and 0.78, respectively, relative to a genome free of TE Therefore, if Juan elements were introduced into a naïve genome, there is a

potential for these elements to reduce the fitness of the host. The loss of fitness would be more evident in a small genome compared to a bigger genome.

Ae. aegypti and *Cx. quinquefasciatus* are globally distributed (Nene *et al*, 2007; Arensburger *et al*, 2010) whereas *An. gambiae* is only found in central Africa (Kiszewksi *et al*, 2004) while *An. darlingi* is only found in South America (Kiszewksi *et al*, 2004; Marinotti *et al*, 2013). In order to adapt to new habitats, *Aedes* and *Culex* would require genomic changes to adapt to new environment (such as changes in host seeking capabilities). A change was observed in the number of odorant binding protein genes, important for the mosquito olfactory system (Manoharan *et al*, 2013). The gene repertoire of odorant binding proteins in *Aedes* and *Culex* have expanded compared to *Anopheles*. It is tempting to explain this increase as a result of TE activity because TE content increased in the genomes of *Aedes* and *Culex*. However, there is no direct evidence showing the increase in TE content led to an increase in this gene family. Manoharan *et al* did not explain how the increase in the olfactory genes occurred. As yet there is no evidence of a direct link between the gene expansion and TE activity- but it might be a fruitful exercise to map TE locations around these loci for evidence of their involvement, e.g as seen in exon shuffling described earlier in Section 1.3.

An initial difficulty was encountered when trying to obtain the full length DNA sequence of JuanA from whole genomic DNA. Despite being present in multiple copies, PCR reactions based on a series of overlapping primers produced multiple bands at unexpected sizes. An explanation of this result is the presence of more 5′ JuanA ends than 3′ ends (Biedler and Tu, 2007). An interesting feature of JuanA is the presence of more 3′ truncations than 5′ truncations. Other retroposons are 5′ truncated due to incomplete reverse transcription. Therefore, if the 5′ primers annealed to a truncated JuanA copies, the excess 3′ primers could have misprimed to genomic DNA. Hence, the PCR reaction would produce unexpected products.

The original published JuanA (M95171) sequence was based on the sequencing of only two JuanA sequences from *Ae. aegypti* Pacific strain, whereas the analysis performed on this chapter is based on the *Ae. aegypti* Liverpool strain (the strain used in the sequencing project). Therefore, sequence differences were to be expected, and indeed, were found between the published JuanA DNA sequence and the JuanA DNA sequence obtained in this study. Most of the nucleotide changes would not cause a significant change in the amino acid sequence. However, the absence of a guanine nucleotide at position 1859 in the M95171 sequence, which causes a frameshift mutation, results in a different amino acid sequence at the beginning of ORF2 (Fig. 3.7) . As a result, Mouches *et al* (1992) suggested that translation of the second ORF involved either by splicing of precursor mRNA, template shifting or termination suppression. This alternative translation is unlikely because termination suppression is the mechanism used because suppression of termination mainly occurs in retroviral transcripts (Bertram *et al*, 2001). The presence of two stop codons consecutively (TAATAA) at the end of the ORF1 sequence would make it even more difficult for suppression of termination to occur. Moreover, splicing of transposable elements is only observed in Class II elements and not among retroposons. Therefore, translation of ORF2 possibly happens after translation of ORF1. The ribosome that has translated ORF1 either shifts upstream and reinitiates translation of ORF2, or recruits another ribosome for this purpose (Han, 2010). As our sequence of this region is supported by 103 full length JuanA sequences found in the *Aedes* genome, it is likely that Mouches *et al* (1992) sequenced this region incorrectly.

The original published JuanC sequence (M91082) included flanking DNA sequences, which were described as part of the element, but apart from that were no contradictions in the start of reading frames or major characteristics of the element compared to currently available data.

These results highlight the need to update DNA databases as more and more sequencing results become available. DNA sequences obtained in the pre-genomics

era should be revised as the wealth of bioinformatics tools make it easier to obtain a better consensus sequence.

Both Juan elements show a high degree of homology in their ORFs. In particular, they possess a unique triple repeat of a cysteine rich region with a consensus of $CX_2CX_4H\ X_4C$-5aa- $CX_2CX_4H\ X_4C$- 9/10aa- $CX_2CX_3H\ X_6C$. This consensus sequence corresponds to the Jockey zinc-knuckle domain. The function of this region is thought to bind and stabilize the mRNA transcript, as well as chaperoning the transcript back into the nucleus for reverse transcription (Laity *et al*, 2001; Ravin *et al*, 2012; Metcalfe and Casane, 2014).

Retroposons are classified to different clades according to their ORF2, which encodes the reverse transcriptase domain (Eickbush and Malik, 2002). A prevailing view is that retroposons have evolved by swapping and combining different ORF1s with ORF2s (Metcalfe and Casane, 2014). Therefore, a mixture of different motifs is found in the ORF1, while little to no change is detected in the ORF2 within a retroposon clade. However, all Jockey elements have only one type of motif in their ORF1, the triple CCHC motif. Thus, it is possible that a triple CCHC motif with a Jockey type ORF2 is a highly advantageous combination for a retroposon. The Juan elements possess these domains.

A current need in humanity's effort to control vector borne diseases, particularly those carried by mosquitoes, is a tool to study their genome. *Drosophila* geneticists were greatly helped by the discovery of P elements and it revolutionised the field of genetics. The mosquito equivalent of P elements has yet been found but I propose that the Juan elements can become part of the toolkit to explore the mosquito genome. They have the ability to spread to high copy numbers, as demonstrated by their high copy numbers in their respective genomes. The ability to spread to high copy numbers also ensures that plenty of active copies are present in the genome and reduces the likelihood that the elements will be inactivated. Jockey elements are only

found in insect genomes, and as a whole, there is no evidence of horizontal transfer of retroposons between species (Eickbush and Malik, 2002). Thus, this reduces the likelihood that the Juan elements might spread to non-target species if released.

The study and understanding of retroposon biology remain important. The genome content of retroposons is high but it is still unclear how and why they reach such high copy numbers. Furthermore, retroposons could be used as a genomic tool. By having a better knowledge of highly prolific and successful elements, the elements can form the genetic toolkit to further manipulate and probe the genome.

CHAPTER 4

CHARACTERISATION OF *CULEX PIPIENS 1*, PIP1, AN ACTIVE LOW COPY NUMBER RETROPOSON THAT HAS A NOVEL START CODON

4.1 Introduction

Retroposons are Class I transposable elements (TEs) that mobilize via a target primed synthesis reaction to generate and insert a copy of the element's mRNA into the host genome. In laymen's terms, it is a copy-and-paste mechanism (Han, 2010; Chapter 1 this thesis). Autonomous retroposons have an intact open reading frame (ORF) coding for the enzyme reverse transcriptase and an endonuclease. Most retroposons also have an additional ORF which codes for a nucleic acid binding domain. Another characteristic is an A-rich 3' tail instead of the long terminal repeats found in retrotransposons. As retrotransposons are also copy-and-paste elements and encode a reverse transcriptase domain, retroposons are commonly referred to as non-long terminal repeat (non-LTR) retrotransposons. A third alternative name is Long Interspersed Nuclear Element (LINE), which originates from the first description of LINE-1 in the human genome. For the purpose of this thesis, the term retroposon will be used exclusively (Eickbush and Malik, 2002; Wicker *et al*, 2007).

Retroposons are transmitted vertically with no strong evidence of horizontal transfer (Eickbush and Malik, 2002), as in the case of retrotransposons or transposons (Maruyama and Hartl, 1991; Robertson, 1993). While there is still considerable paucity of data, retroposons found in one host species do not always appear to have relatives in a closely related host species. This does not contradict vertical inheritance, but rather reflects the rapid dynamics of gain and loss of elements in relation to their

activity and impact on the host. There is also scope for modular evolution, where a retroposon can be formed *de novo* by exchange of genetic material within the genome. For example, an intact RT and endonuclease domain could be transcribed downstream next to an unrelated functional ORF1 domain. This element could then mobilise and hence evolves within the genome to become so different from its original components that it is no longer recognised as being related to its ancestral sequences, but is a novel sequence.

In addition, retroposons display mobilisation using a master gene. The master gene hypothesis states that only one gene locus is used to generate new copies (Kass *et al* 1995). In the case of TEs, the copies of the master gene can themselves become a master copy gene, thus generating even more copies in the genome. As all of the copies are related, a phylogenetic tree can indicate if this is true. Copies generated from master copy genes would be very closely related and have only a single origin. If the copies themselves can function as a master gene, the phylogenetic tree would have multiple branches within each node (Johnson and Brookfied, 2006).

Identification of retroposons has been greatly aided by the advent of whole genome sequencing and bioinformatics tools. Prior to this, identification of elements greatly relied on the conventional molecular biology. Potential new elements have to be isolated, cloned and sequenced. This is a tedious and laborious process. Moreover, TEs are present in multiple copies and it is difficult to identify the most abundant copy using this method. Now, a researcher can mine whole genomic data using a vast array of bioinformatics tools to identify TEs (Durand *et al*, 2006; Janicki *et al*, 2011).

Here, I describe a retroposon found only in *Culex. pipiens* s.l., termed Pip1. I use both experimental molecular biology and bioinformatics to comprehensively characterise the element. Pip1 elements were originally identified by Crainey *et al* (2005) and are grouped in the Jockey clade. Pip1 elements show signs of being recently active, but in stark contrast to the Juan elements studied in the previous chapter, they are present in low copy numbers, despite being in a large, low gene density genome

(Crainey and Malcolm, 2010). As a central theme to the overall project is the hypothesis that low gene density genomes are permissive for unrestricted transposition, it was of interest to look for clues as to why Pip1 has apparently not been as successful as Juan.

Since the earlier studies, the data from the *Cx. quinquefasciatus* genome sequencing project (Arensburger *et al*, 2010) and subsequent projects on other members of the *Cx. pipiens* s.l. complex have become available to allow a more comprehensive survey of Pip1. Here this has been used to examine possible explanations for the relatively low copy number of Pip1, which may include regulation of transposition, limited capacity for transposition, or a recent origin. Intact elements and easily identified truncated copies that show reoccurring truncation patterns, point to a recent origin, but other characteristics including sub-groups and a missing or unusual start codon suggest something less simple.

4.2 Materials and Methods

4.2.1 Mosquito DNA extraction

Refer to Section 2.1.6

4.2.2 PCR

To amplify Pip1, primers flanking the elements were designed using Primer-BLAST (Rozen and Skaletsky, 2000). Expand™ Long Template PCR system (Roche, Germany) was the PCR system used to obtain the full length element. Each PCR reaction contained 10μM of primers of forward primer, Pip1 Fla 1745F (AAATCGACTCTCGTGTTTGGA), and reverse primer, Pip1 Fla 6243R (GCTCCAGGATGTTACATTTGC); 500μM of dNTPs; 5μl of 10x Expand Long Template Buffer 1 (17.5mM $MgCl_2$); 0.5 units of Expand™ Long Template Enzyme mix. The PCR programme was (1) Heat PCR machine to 120°C. (2) 93°C for 2 minutes.

(3) 10 cycles of 93°C for 10 seconds, 50°C of for 30 seconds, and 68°C for 4 minutes (4) 25 cycles of 93°C for 15 seconds, 50°C for 30 seconds, and 68°C for 4 minutes + 20 second for each successive cycle. (5) Final elongation at 68°C for 7 minutes. PCR products were analysed on an agarose gel.

4.2.3 Bioinformatics

The genome sequences were mined using BLAST (Zhang *et al*, 2000) on the NCBI website, using the reference genomic sequences as database. All parameters were set to their default values.

Alignments and construction of phylogenetic trees were carried out using CLC DNA Workbench Version 6.0.2 (CLC Bio, Denmark). Repeat Masker was run on http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker (Smit *et al*, unpublished). YASS dotblot was performed to find sequence similarities within the Pip1 DNA sequence (Noe and Kucherov, 2005).

Jockey elements were obtained from Repbase (Jurka *et al*, 2005). Only intact autonomous elements were used. Sequence for CM-gag was obtained from Bensaadi-Merchermek *et al* (1997) while the Juan elements were originally obtained from Mouches *et al* (1992), Agarwal *et al* (1993) and from Chapter 3 (this thesis).

4.3 Results

4.3.1 Pip1 sequence and key structures

Pip1 was originally identified in a bioinformatic screen by Crainey *et al* (2005) and this sequence, Pip1 3.19, is used as the reference sequence (The number 3.19 refers to the contig where the element is found in the genome sequencing project). To validate the bioinformatics result, the element was cloned in the laboratory and sent for DNA sequencing.

The DNA sequence is presented in Figure 4.1, and the important features are highlighted. Pip1 is at 4387bp long and is close to the average for Jockey clade elements. The TTCG box, present at position 2, is also typical of Jockey promoters. There are two long open reading frames (ORFs) and they overlap for 13 nucleotides. The element terminates at the 3' end with four repeats of TTGAA. The AATAAAA polyadenylation signal precedes the repeats.

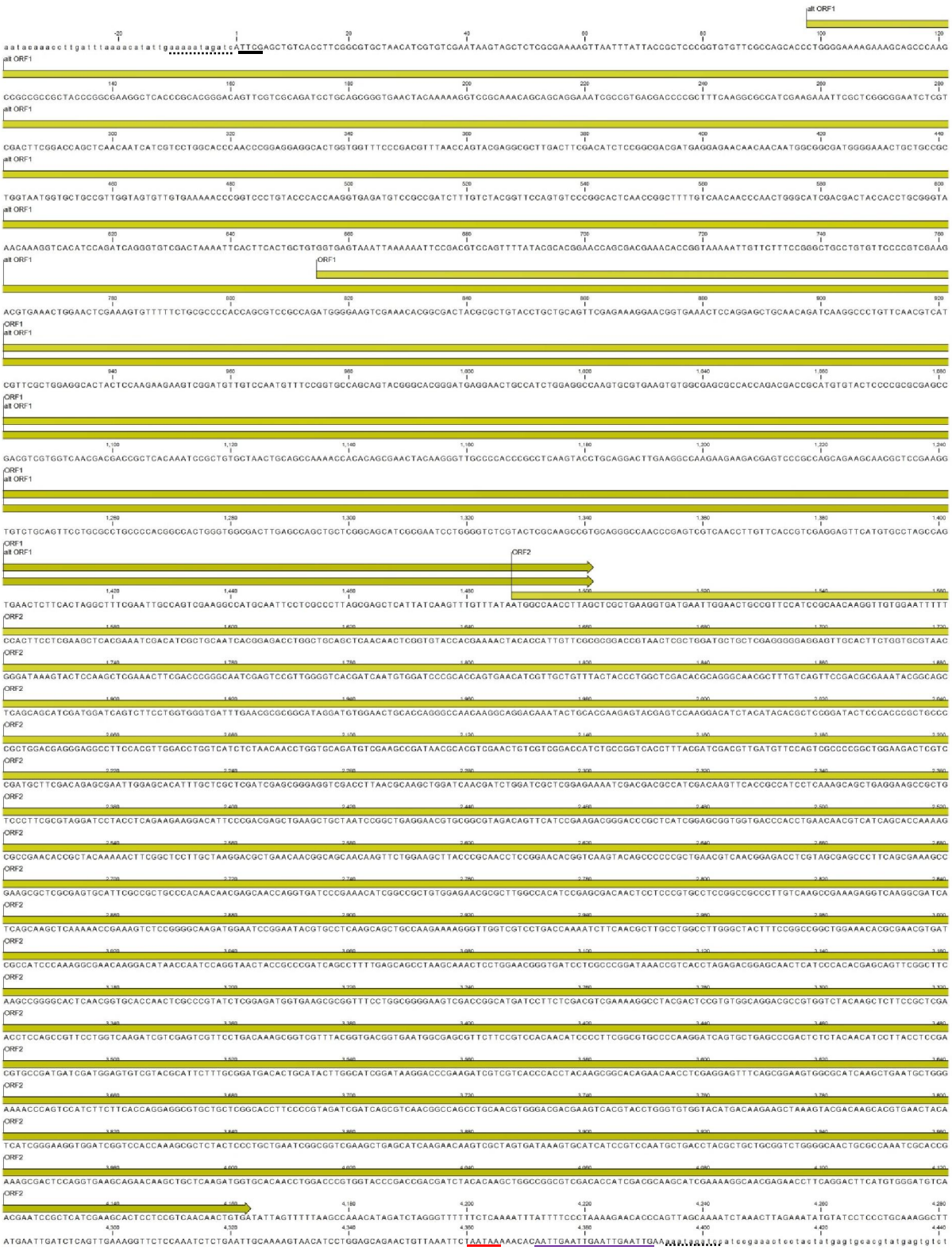

Figure 4.1 Nucleotide sequence of Pip1 3.19. Pip1 DNA sequence is shown in upper case letters while flanking regions are shown in lower case letters. Target-site duplications are shown with a dashed line. The TTCG promoter sequence characteristic of Jockey elements is underlined. The ORFs are annotated in the diagram, including the putative longer ORF1 (alt ORF1). The AATAAA polyadenylation signal is underlined (red line) while the AATTG repeats at the 3′ end is shown with a purple line.

4.3.2 Polymorphism in Pip1 insertion sites

The PCR to obtain the full length element was initially performed on *Cx. quinquefasciatus* Muheza and TRR1 strains. However, instead of the expected band size at roughly 5kb, a smaller band size, around 550bp, was obtained (Figure 4.2). The smaller band size was analysed and sent for DNA sequencing. The DNA sequence results matched the sequence of genomic DNA without the Pip1 insertion.

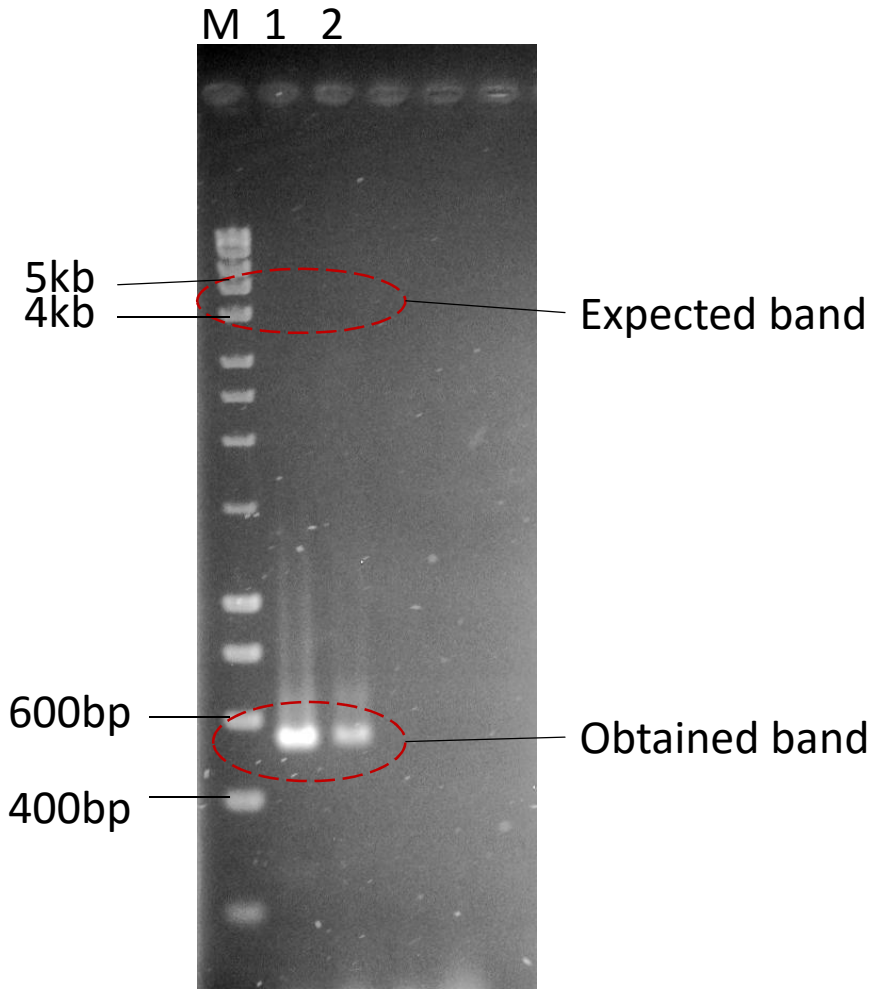

Figure 4.2 Gel electrophoresis result of PCR on *Culex quinquefasciatus* Muheza (1) and TRR1 (2) strain using flanking primers. The product about 550 bp was obtained rather than the expected product at 4.9kb. Expand Long Template PCR system was used. M is the Bioline HyperLadder 1 (Bioline, UK). The 1.0% agarose gel was run at 90V for 4 hours.

In addition, the initial PCR performed on *Cx. quinquefasciatus* whole genomic DNA from the Johannesburg strain also gave a mixed result, as shown in Figure 4.3. In one of the lanes, both band sizes were obtained. The smaller band size without the 5kb fragment was obtained in 2 of the other PCR reactions.

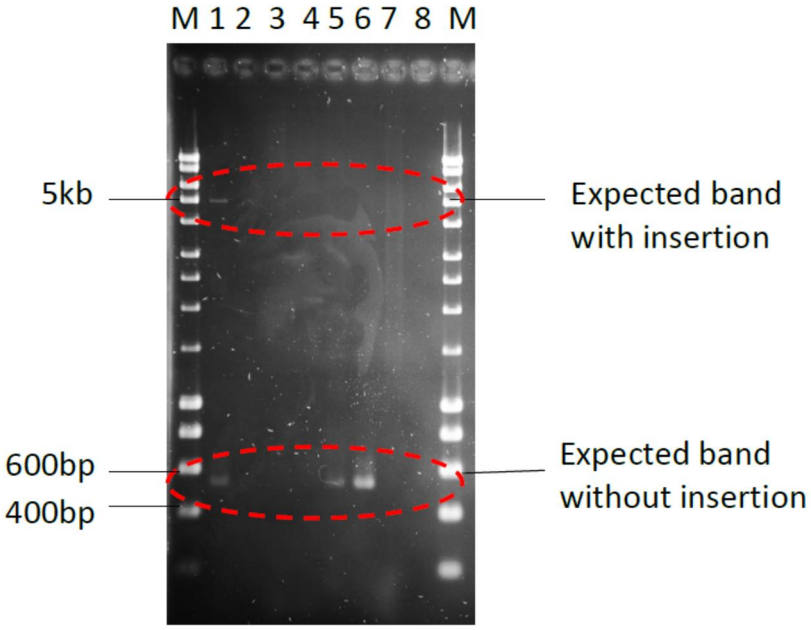

Figure 4.3 Gel electrophoresis result of PCR on *Culex quinquefasciatus* Johannesburg strain. In lane 1, 2 bands were obtained- the expected band with insertion at 5kb, and another one at 550bp. In the other PCRs, none or only the 550bp band was obtained. Expand Long Template PCR system was used. M is the Bioline HyperLadder 1 (Bioline, UK). The 1.0% agarose gel was run at 90V for 4 hours. The different lanes represent different individual *Cx. quinquefasciatus* DNA.

4.3.3 Full length or near-full length Pip1

The full length sequence was run on NCBI using BLAST tool against the whole genome sequencing database (Zhang *et al*, 2000; Morgulis *et al*, 2008). All hits were to sequences from *Culex pipiens* s.l. (data not shown). Twenty three full or near full length Pip1 copies was identified (Table 4.1 and Fig 4.4). In order to distinguish each copy, each Pip1 copy is named after the number of the supercontig. Each copy was checked for the presence of the 5′ promoter, the number of 3′ tail repeats and if they still maintained coding potential for in the ORF1 and ORF2. The result is presented in Table 4.1. There were 13 putative intact full length Pip1 copies.

Table 4.1 Summary of the 23 full length or near full length Pip1 copies in the genome. ✓ indicates an intact reading frame , × indicates the reading frame is no longer intact, * sequence identity with 3.19, - space

Contig	Identity (%)	5' (promoter)	3' end	ORF1 (amino acid length)	Intact ORF2
3.19	100	CATTCGAGCTGT	TTGAATTGAATTGAA	228	✓
3.244	99	G**********	**T*T**A*T*AATN	228	✓
3.208	99	**********	**----CT****CT*	228	✓
3.246	99	**********	*****AATTTCAAGT	228	✓
3.679	91	**********	*****AATTGAAA*C	228	×
3.2223	99	TTAAG*******	**********	228	×
3.538	99	**********	**********	×	×
3.444	94	**********	**********	381	✓
3.15	94	**********	*****A*CT****T*	381	✓
3.47	93	**********	No tail (-89 bp)	381	✓
3.1861	91	G**********	***G***A*******	381	✓
3.352	93	**********	*****A*TG*AAA*T	381	×
3.122	93	G**********	*****A--C******	381	×
3.1071	92	G**********	*********---**T	428	✓
3.33	92	**********	**********	428	✓
3.10	92	**********	**********	428	✓
3.185	91	GTA*********	**********A*AT*	428	✓
3.147	91	**********	No tail (-179 bp)	428	✓
3.162	92	AG*AT*******	No tail (-147 bp)	428	×
3.34	89	TT**********	**********A*AT*	428	×
3.251	87	**********	**---C*A****CT*	428	×
3.1149	91	**********	**********	×	×
3.1249	91	**********	**********	×	×

4.3.4 Palindromic sequence within Pip1

The vast majority of Pip1 elements identified by the Pip1 3.19 BLASTN search are 5′ truncated; of which 104 have an intact 3′ terminal and 34 do not. Some of the latter are missing only a small portion of the 3′ tail, but nevertheless this sub-group are likely to have been disrupted by large insertions or recombination. That might also be the cause of some of the 5′ truncations, but the observation of so many is consistent with the conventional model that reverse transcription frequently terminates early during the insertion of the element. This was investigated further by comparing the distribution of 5′ truncation points for elements with intact 3′ termini, so essentially the length of each element, but using a common 3′ starting point based on the alignment with Pip1 3.19 (Figure 4.4).

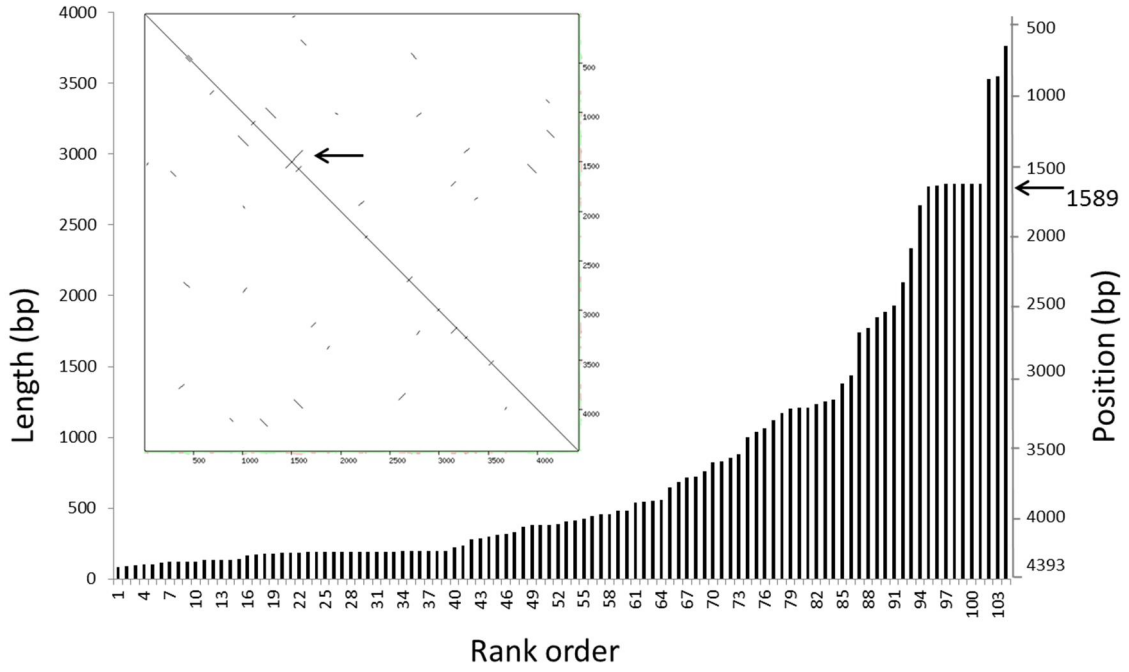

Figure 4.4 Distribution of length variation amongst 5′ truncated Pip 1 elements. The elements are distributed in rank order according by length from 3′ to 5′ end of 5′ truncated Pip 1 elements with an intact 3′ terminal. This is based on a BLASTN search with query Pip1 3.19 (4387 bp) against the Reference Genomic Sequences (refseq_genomic) database, using the megablast setting for highly similar sequences. The right hand Y axis indicates equivalent residue positions (5′ to 3′) in the sequence for Pip1 3.19 (Figure 4.1). The arrow 1589 indicates the truncation position for four elements. The insert is a screenshot of a YASS dotblot of the Pip1 3.19 DNA sequence aligned against it-self. The arrow indicates the highest scoring internal alignment.

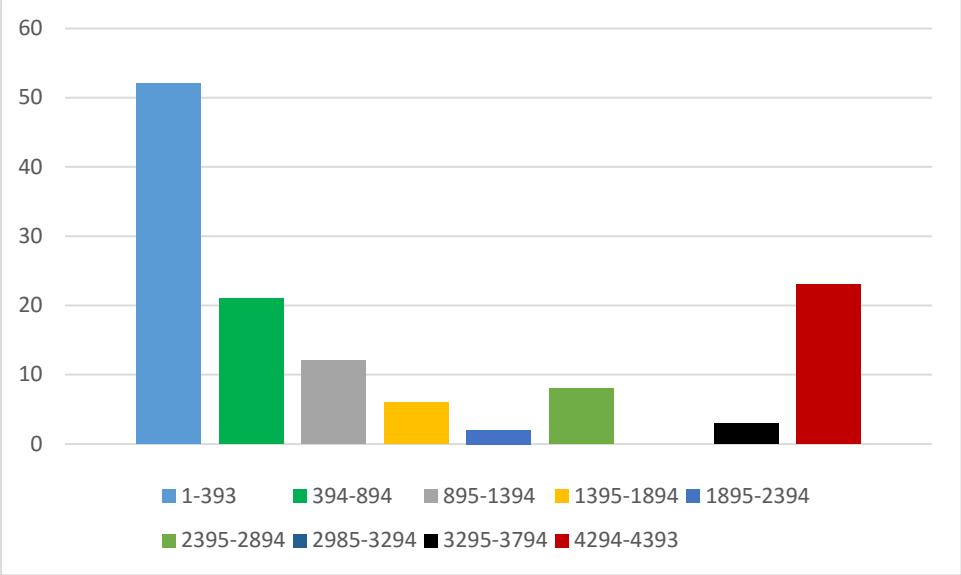

Figure 4.5 Histogram showing the frequency distribution of Pip1 elements. The elements are grouped based on their length from the 3′ end of Pip1.

Figure 4.5 shows the distribution of lengths of Pip1 copies. Apart from the full-length elements (in red), the truncated elements show an approximately exponential distribution, which is expected from a random uniform frequency of termination. This pattern can be explained as a consequence of the mobilisation process. The reverse transcriptase transcribes the element beginning from the 3′ end. However, the transcription can be incomplete (a possible reason being the enzymatic machinery falling from the DNA (Han, 2010)), generating copies with only the 3′ end Pip1. If this termination occurred uniformly and randomly during reverse transcription, the distribution of lengths would be exponential. However, close inspection reveals some deviations from the exponential curve (Fig. 4.4), most obviously in the large number of transcripts of length 2804bp. This can be explained by the palindromic sequence in the region. A dotplot alignment of the Pip1 3.19 sequence against itself (Figure 4.4 insert) shows several points where the reverse of a sequence gives a significant alignment to the forward sequence indicative of a complete or partial palindrome. These are displayed as short lines crossing the line of identity at right angles to it. The longest one, and highest scoring of the internal alignments (score = 139, bitscore =

42.81) (arrow in Figure 4.4 insert) coincides with the truncation hotspot at position 1589.

The capacity for this sequence, which extends from position 1370 to 1598, to fold back on itself and form a hairpin loop is illustrated in Figure 4.6. There is about 57% complementarity within the loop, with the truncation point occurring close to the 3′ end. These observations are entirely consistent with 5′ truncation occurring because the reverse transcriptase failed to progress past the loop. It is notable that this partial palindrome exactly encompasses the overlap between ORF1 and ORF2 (Figure 4.1).

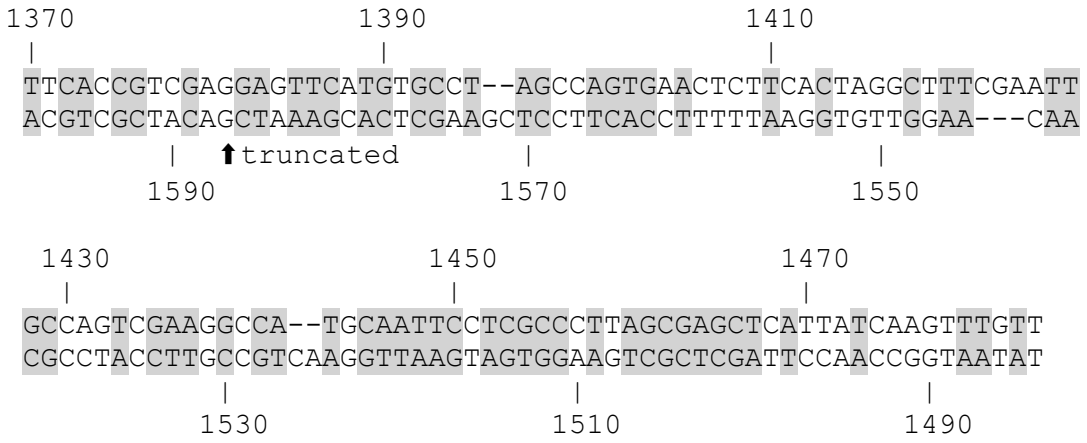

Figure 4.6 A palindromic sequence corresponding to a Pip1 5′ truncation hotspot. An alignment of Pip1 3.19 sequence positions 1370-1483 (Figure 4.1) with the reversed sequence from positions 1484-1598 to show complementarity (shaded in grey). The arrow indicates a truncation hotspot; all sequence from the residue indicated towards the 5′ (decreasing position number) would be missing.

Flanking regions immediately next to the full length Pip1 copies were analysed for target site duplications and evidence of sequence specificity (Table 4.2). Good candidates for TSDs ranging in size from 8 to 17 bp were found for 15 elements. No evidence for TSDs were found for the three elements that were 3′ truncated (3.162, 3.47, 3.147) and sequence data is not available for the 3′ flank of 3.244 (see Table 4.1), so these were not included in Table 4.2. Element 3.1149 was not flanked with candidate TSDs longer than 6 bp, however the 3′ flanking sequence is identical to that for element 3.1249, which does have a convincing TSD. Similarly, the 3′ flanking

regions of elements 3.185 and 3.34 are identical and only 3.34 has a TSD. There is almost sequence identity along the 3′ half of each pair of elements, that contrasts with multiple substitutions along the 5′ half and entirely different sequences in the 5′ flanks. It is likely that 3.1149 and 3.185 are assembly artefacts or the products of recombination. There were no indications that that is true for 3.251, or any other obvious explanation for why no TSD was found associated with this element. It is notable that it does not have a distinct 3′ tail with GAATT, but then that is also true of 3.208. Despite having intact ORFs, 3.251 does contain a long substitution (568 bp) close to the 3′ terminal, indicating that it has been disrupted by recombination.

Table 4.2 Target site duplications in Pip1 5′ and 3′ flanking regions. Putative target site duplications (TSDs) are underlined and highlighted in grey. The length of the TSD is indicated in the final column, X indicates that no good TSD candidate was found. Sequences are aligned relative to position numbers for 3.19 (Figure 4.1)

5' flanking regions (to position 4)	contig	3' flanking regions (from position 4377)	bp
CGACTAAAAACCATTTTGGTCACATTC	3.444	ATTGAAAAAAACCATTTCTGATCACTTTTTGTCA	11
ATGTATAAAAATAAAAAAAAATCATTC	3.352	AATTGAAAAATAAAAAAAATAATGAAAAAAATAA	14
TTTTTTTTAAATTAGAATTTTTCATTC	3.15	AATCTATTGTAAATTAGAATTTACAAAGTTAGAT	10
GCTGTAAGAATATCCAGCTCTGGATTC	3.122	AACATTGAAACATATCCAGCTCTGTGAGAACTCT	13
ACCAATAAAATATAATTAATCTCATTC	3.33	ATTGAATTGAATAAAATGTAGTACCTTGTTCTAC	8
CTGAGTTCAACAACCCACTTTTCATTC	3.679	AAATTGAAAAACAACCCACTTTTTCATACGAATT	14
ATTTTTTATATACGAAAACTTTTTTTC	3.34	AGAAATATACGAAAACAAAATTCATTAATTCTAT	12
TCGTTGGCAAGAGAAAAATAATGTATC	3.185	AGAAATATACGAAAACAAAATTCATTAATTCTAT	X
TAAGAATTTAAGAATTTAAGAAGATTC	3.1071	ATTGAGATTTAAGAATTTAAGAATTTAAGAATTT	17
TAAATTCTTAAATTCTTAAATTCATTC	3.1149	ATTGAATTGAATTAAATTGAAGAATTTCAGAGTT	X
TGAGAATTTATAAATTGAAGAACATTC	3.1249	ATTGAATTGAATTAAATTGAAGAATTTCAGAGTT	12
AAATTCTTAAATTCTTAAATTCTTAAG	3.2223	ATTGAATTGAACTATTTAAACAACTAATAAACAA	X
AAATTTTCGTAAAAAATGCGATCATTC	3.538	ATTGAATTGAAAATTGAAAAGTAAAAAATGCGAT	14
TAAAACATATTGAAAAATAGATCATTC	3.19	ATTGAATTGAATTGAAAAATAGATCCATCCGAAA	11
AAATTCTTAATTAGACCTTGACCATTC	3.208	AATTCTAATTAATTAGACCTCCTAGACCCACCTT	12
TATTTCAAGTTTTTTTTTTTTTCATTC	3.246	AAATTTCAAGTTTTTCAATACTACCCTACAAAAA	12
AGGAATTAAAAGAAAGTAATGTCATTC	3.251	AAATTCTAAAAAATTGCATGAATGCATCATCCAA	X
AGCTACTTTTTTAACCCAAACTCATTC	3.10	ATTGAATTGAATTAACCCAAACTCAGAAAAATTG	13
AAAGAAAGAAAATGCAACAAAGGATTC	3.1861	ATTAAATTGAATTGAAAATGCAACAAAAAGCTCA	14

Sequences of 30 nucleotides upstream and downstream of each Pip1 copy were compared by alignment to identify target site specificity. A specific target sequence or

consistent motif was not found, but all of the TSDs are AT rich, with only 3.122 close to 50%.

4.3.5 Analysis of ORF2: Pip1 is confined within the *Culex* genome and is a Jockey element

The inferred sequence of amino acid residues from ORF2 was determined for each of the 13 elements in which ORF2 was intact and then aligned. A phylogenetic tree based on the alignment (Figure 4.7) produced three distinct monophyletic groups (shown as A, B and C in the figure). This is consistent with observations on alignments of the nucleotide sequences, which showed evidence of sub-groups. However, attempts to resolve the groups proved difficult, because only certain blocks appeared to sub-divide whereas others were uniform. Furthermore subdivision within one block of the alignment did not necessarily agree with another. It was decided to simplify the problem by focusing first on the inferred ORF2 protein. This, and in particular the reverse transcriptase domain, is the most conserved part of the element and is traditionally used to classify retroposons.

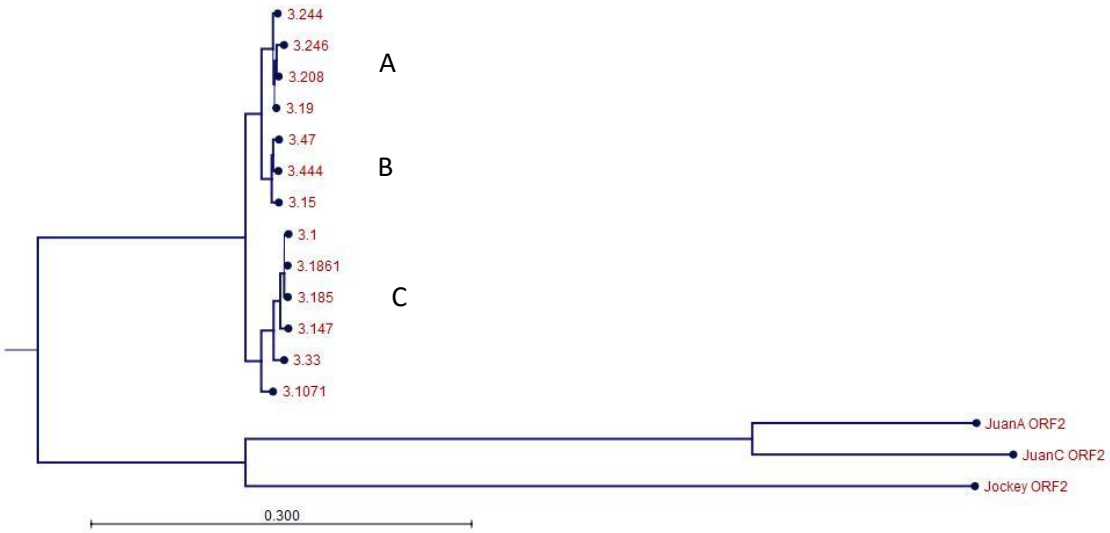

Figure 4.7 Phylogram constructed from the ORF2 of all 13 intact Pip1 copies. The elements separate into three different groups, marked as A, B and C in the diagram. The Juan and Jockey elements are set as outgroups. Bootstrap values more than 80% are indicated with a darker branch line.

A phylogenetic tree was constructed using the inferred ORF2 amino acid residue sequences from Pip 1 and from other Jockey elements taken from RepBase. The result is presented in Figure 4.8. Bootstrap values for the majority of the tree is more than 80%, thus, providing a high support for the phylogeny constructed. The results agree with other studies done on retroposon phylogeny (Eickbush and Malik, 2002; Crainey *et al*, 2005; Metcalfe and Casane, 2014).

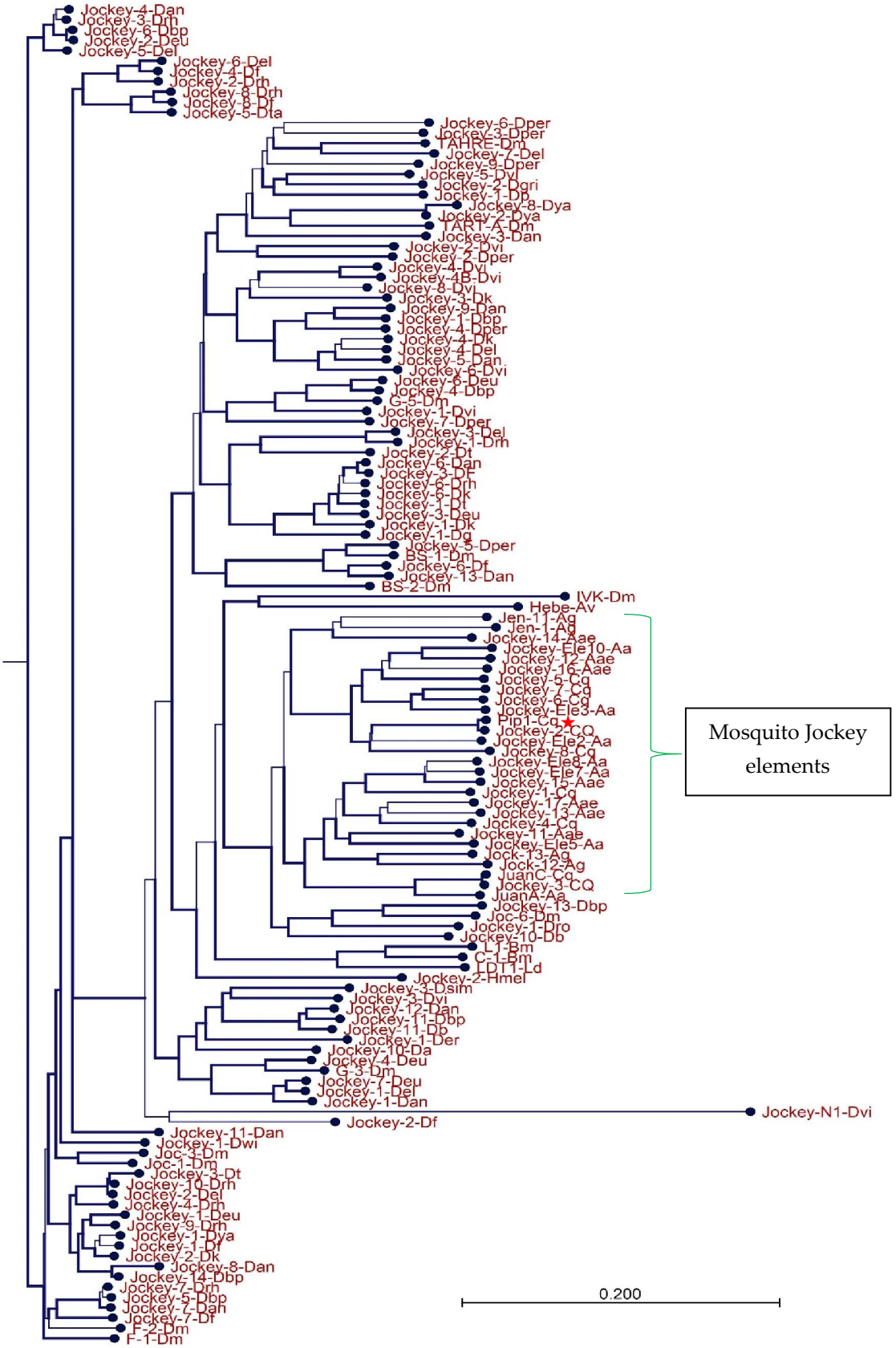

Figure 4.8 Phylogenetic tree constructed with the ORF2 of Jockey elements obtained from Repbase as well as the Pip1 3.19. The Pip1 element is marked with a ★. All Jockey elements from mosquitoes are grouped together from a single branch root, as indicated. Bootstrap values more than 80% are indicated with a darker branch line.

All of the elements from mosquitoes came from a single branch root (indicated in the figure). The sister group to the mosquito Jockey elements appear be Jockey elements from fruitflies, including *D. melanogaster*.

The Pip1 element (marked with a ★) is located next to Jockey-2 element from *Cx. quinquefasciatus*, which was generated from a consensus of 10 copies with >99% identity. This Jockey element is likely to be Pip1. The results confirm the placement of Pip1 in the Jockey clade (Crainey *et al*, 2005) and despite evidence of variant Pip1 elements; these belong to sub-groups not to different families of element.

4.3.6 The ORF1: Alternative start codons, CCHC zinc finger and similarity to other elements

The start of the ORF1 of Pip1 is harder to define. Pip1 3.19 has a potential ORF1 of 687bp, coding for a 228 amino acid product. In comparison to the ORF1 of other Jockey retroposons, this sequence is shorter than expected. The longest ORF1 product with an ATG start codon is found in elements in Group C, which at 1287bp and coding potential for a product of 428 amino acids long. Elements in Group B and A can encode a product of 381 and 228 amino acids respectively. A longer ORF can be predicted if an alternative start codon, CTG for leucine, (Touriol *et al*, 2003) is used. This codon is present in all of the groups and would produce a 471 amino acid product. The phylogram of intact Pip1 ORF1 also matched the phylogram constructed from using Pip1 ORF2 (Figure 4.9).

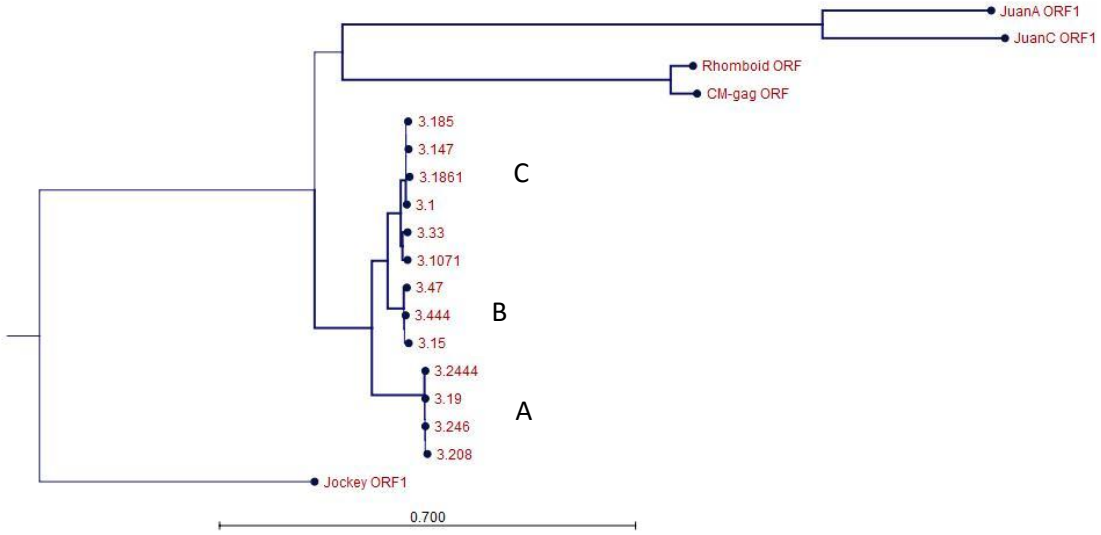

Figure 4.9 Phylogram constructed from the putative 471 amino acid product of ORF1 of all 13 intact Pip1 copies. The elements separate into three different groups, marked as A, B and C in the diagram. The Juan and Jockey elements are set as outgroups. Bootstrap values more than 80% are indicated with a darker branch line.

The phylogeny for groups A, B and C is reversed when comparing the ORF1 tree (Fig. 4.9) with the ORF2 tree (Fig. 4.7). This could be due to the inclusion of Rhomboid ORF and the CM-gag ORF in the ORF1 tree. The outgroups are in a different branch while the Pip1 copies form a monophyletic group. The bootstrap values for the lineage to Jockey ORF and the lineage between the Pip1 copies with the other elements are also less supported (bootstrap <80%).

The CCHC zinc finger is found in Pip1. The motif is repeated three times and is characteristic of zinc fingers found in Jockey elements. As this sequence is present at the 3'end, this motif is intact in the three separate groups. By running a BLASTx on the ORF1, it recovered two hits with a high score, namely the 3'end of Rhomboid but more interestingly to CM-gag (38% identity). CM-gag is another transposable element found in the *Cx. quinquefasciatus* genome (Bensaadi-Merchermek *et al*, 1997) while Rhomboid is a transmembrane protease (Figure 4.9). CM-gag is a unique mobile element in the sense that it only has a ORF1 and does not possess a ORF2, and it is estimated that the *Culex pipiens* genome has 150 copies of CM-gag. By running the Rhomboid DNA sequence on GeneValidator (Dragan *et al*, 2014), the sequence was

validated as a combination of two genes. CM-gag could have inserted at the 3′ end of the rhomboid; however, the amino acid sequence of rhomboid is definitely annotated incorrectly.

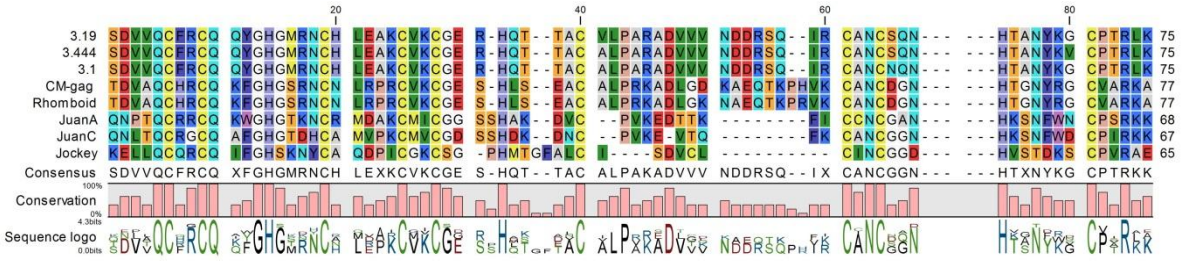

Figure 4.10 Alignment of the CCHC zinc-finger motif. Conservation of the CCHC motif is found in the three separate groups. Three Pip1 copies, each from a separate group, are shown, together with CM-gag, Rhomboid, the Juan elements and Jockey element. The triple repeat of CCHC is consistent with the motif found in elements within the Jockey clade.

Using the ORF1 from Pip1 3.19 and other Jockey elements, a phylogram was constructed (Figure 4.11). The Pip1 element (marked with a ★) is within the same branch node to CM-gag (marked with a *). The tree is different to the phylogeny based on the ORF2 (Fig. 4.8) and the mosquito Jockey elements are not monophyletic. In addition, most of the nodes of the tree do not have bootstrap values more than 80%. A possible reason for the low bootstrap value is recombination between the elements. The ORF1 from an element could have been transferred to another different transposable element (Metcalfe and Casane, 2014). This possible explanation is consistent with previous findings that the ORF1 could be exchanged between different lineages and is not an appropriate sequence for phylogeny reconstruction (Eickbush and Malik, 2002). Therefore, phylogeny reconstruction of retroposons are based on the ORF2 sequences.

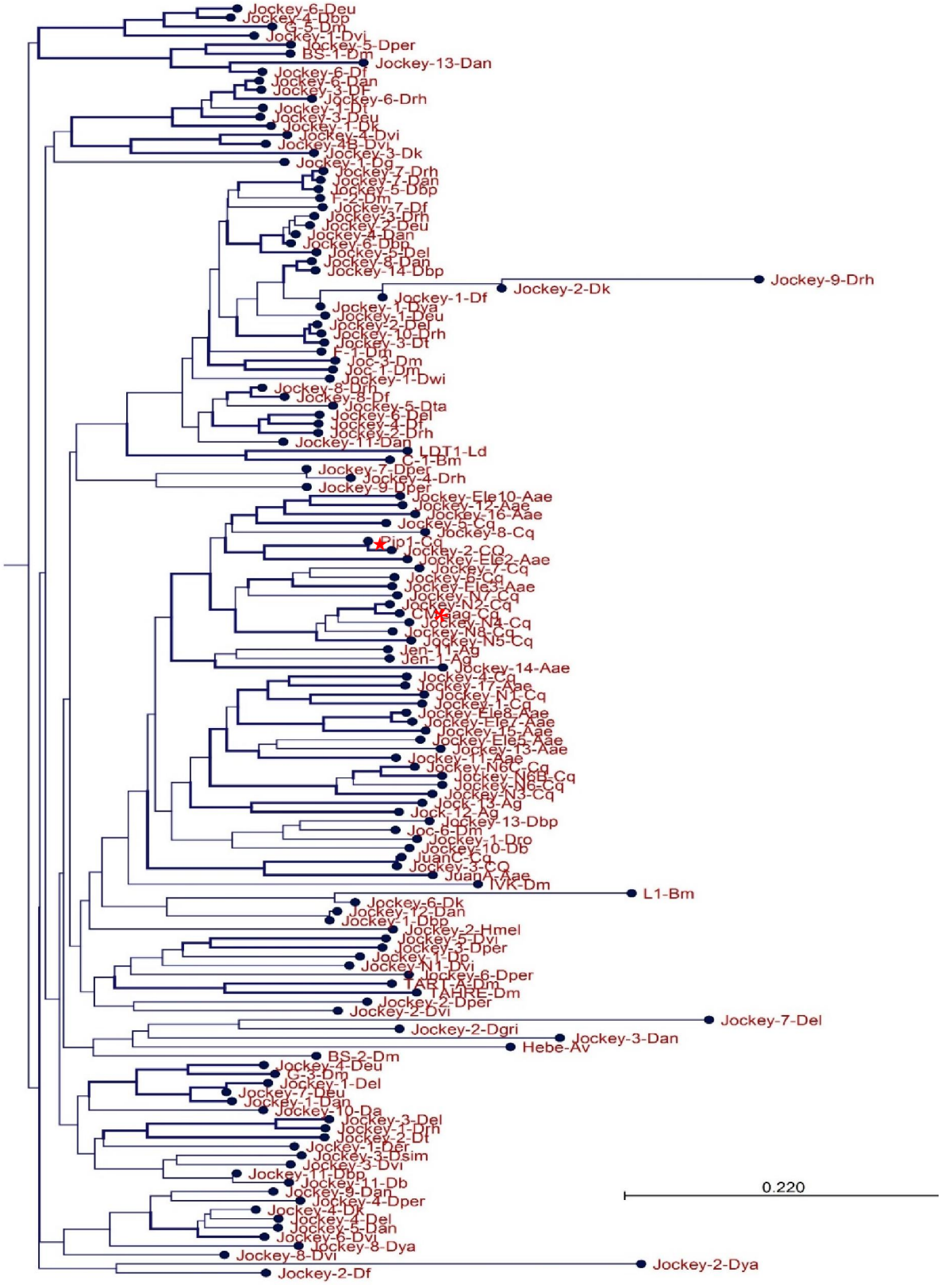

Figure 4.11 Phylogenetic tree constructed with the ORF1 of Jockey elements, Pip1 and CM-gag. The Pip1 element is marked with a ★ ; Cm-gag is marked with a *. Bootstrap values more than 80% are indicated with a darker branch line.

4.4 Discussion

Pip1 is a retroposon roughly 4387bp long and has no long terminal repeats. It is found in the *Culex* genome and only has 13 full length intact copies. The low copy number could be due to a recent de novo origin of Pip1 elements. It seems most likely that because of its recent 'birth', Pip1 simply has not yet had enough evolutionary time to reach a high copy number. In addition, some of the insertion sites from the Johannesburg strain are not present in the Muheza or TRRI strains. This suggests that Pip1 is polymorphic between strains and the insertion occurred after the strains diverged from each other. This interpretation is supported by a previous report of polymorphism (Crainey and Malcolm, 2010). In addition, Pip1 elements can be identified in other *Cx. quinquefasciatus* strains. The data is more consistent with a Pip1 element being present before the strains diverged, which remained active after the divergence.

Analysis of the ORF2 shows that Pip1 fits the master copy gene hypothesis to some extent, using multiple master copy genes (Figure 4.7). The phylogenetic tree shows the intact copies grouping into three distinct groups. Group A copies likely arose from a single master copy gene. Copies from Group B and C are more closely related to each other, suggesting another master gene generated these Pip1 copies. Another possibility is that Pip1 undergoes several bouts of transposition: the periodicity observed might be due to a full length copy being successfully generated.

Regardless, Pip1 mobilisation generates a lot of truncated Pip1 copies. 10 inactivated full length copies in addition to multiple truncated copies were identified in the genome. Generation of the multiple truncated relic copies at 1589bp were due to secondary structures as shown in Figure 4.6. The hairpin loop could have stopped reverse transcriptase activity and prematurely terminating reverse transcription. As there are many insertion events, it does not seem that Pip1 is regulated by a specific mechanism.

When other Jockey elements are included in the ORF2 analysis, it is clear that Pip1 is strictly in the Jockey clade (Figure 4.8). It is only found within *Culex*; thus it displays strict vertical inheritance and there is no evidence that it arose from horizontal transmission. By analysing only the 13 intact copies, the copies group into 3 distinct groups with different branch lengths. There is a high genetic difference between the copies.

Pip1 could also be using an alternative start codon for translation of its ORF1. The major difference of Pip1 copies is the length of the longest ORF1 detected using a methionine initiation codon. However, a potential alternative start codon (leucine) is present in all of the copies, very early in the sequence (Touriol *et al*, 2003), and could be potentially used to start translation.

While the ORF1 region of retroposons is not highly reliable for phylogenetic reconstruction, the fact that Pip1 ORF1 shares homology (38% identity) with CM-gag, another repeat element but only with a gag-like protein is intriguing. It might be possible that a retroposon could have inserted downstream of an ancestral version of CM-gag copying an intact ORF2 but truncating at the beginning of ORF1. This chimeric element, containing CM-gag with this ORF2 could then have mobilised and eventually given rise to Pip1 elements. The converse could also have happened: an ancestral version of CM-gag transposed directly upstream of an ORF2 of a retroposon.

Pip1 is present in the relatively large genome of *Cx. quinquesfasciatus* (579Mbp), and 29% of the genome consists of TEs. Interestingly, within this genome, JuanC (Agarwal, 1993; chapter 3 in this thesis) is present in high copy numbers. However, both Pip1 and JuanC are Jockey elements. However, Pip1 only has 13 full length active copies while JuanC has potentially 2500 active copies. There are a few possibilities why Pip1 copy number is less than JuanC. Pip1 could be a newly evolved element and might reach a high copy number given enough evolutionary time. This is evident in

the strain polymorphism observed between *Culex pipiens* strains. Pip1 is still actively transposing.

Alternatively, the host might regulate Pip1 activity and restrict the increase in copy number, although this is unlikely since JuanC does not appear to be regulated and have achieved a very high abundance. The polymorphism displayed by Pip1 also indicate that Pip1 is actively mobilising without being completely restricted by the host.

A more likely explanation is the tendency for Pip1 to mobilise incompletely. Only 13 full length elements were identified, another 10 full length copies are inactivated. In addition to full length elements, multiple truncated copies are present (Fig. 4.4). Pip1 does not mobilise completely and generates truncated copies. This prevents Pip1 from reaching a high copy number in the genome.

CHAPTER 5

ARTIFICIAL HORIZONTAL TRANSFER OF RETROPOSONS

5.1 Introduction

Since Thomas Morgan's pioneering work, *Drosophila melanogaster* has become the model organism in various areas of research, including genetics. In addition to its short generation time, it is also easy to keep and maintain in the lab. Consequently, an array of techniques and tools has also been developed to aid in using *D. melanogaster* as a model organism, including protocols for germline transformation – the approach used in this chapter.

Germline transformation is the introduction of DNA into the germ cells of a different organism. As *D. melanogaster* embryos are small and can be easily manipulated, they are suitable for microinjections. In the early stages of development, the *D. melanogaster* embryo is a syncytial blastoderm (containing multiple nuclei not separated by membranes), making it easier to incorporate foreign DNA.

Many different gene vectors have been used to introduce foreign DNA into *D. melanogaster*. piggyBac is a transposon originally identified in the cabbage looper moth, *Trichoplusia ni* (Cary *et al*, 1989; Fraser *et al*, 1995). It has been widely used and developed as a tool for *D. melanogaster* germline transformation. Transposition of piggyBac only requires a functional transposase and the terminal inverted repeats. A vector is created by inserting a gene of interest between the terminal inverted repeats and removing most of the intervening sequences; the transposase can then be supplied in *trans* to affect transposition of the gene. This method has an added benefit of

ensuring that piggyBac cannot be mobilised once inserted because it lacks a functional transposase.

Retroposons are Class I transposable elements. They mobilise using reverse transcription of mRNA copies. Since RNA is very much less stable than DNA it is unlikely to survive outside a living organism – which may reduce the probability of horizontal transmission between species (Eickbush and Malik, 2002); indeed no incontrovertible evidence of horizontal transfer has been found. Rather, inheritance of retroposons is thought to be strictly via vertical transmission only. If horizontal transfer were to occur the impact of a newly arrived element is difficult predict, but conceivably it would be similar to observations made on P elements, which are the Class II transposons.

P elements causes a syndrome of sterility, mutations and increased recombination called hybrid dysgenesis in the offspring of *D. melanogaster*. The effect on the offspring is determined by the cytoplasmic contents, or cytotype, of the maternal fly (Engels 1989). When a female fly with active P elements in the genome (P cytotype) mates with any male, the P elements will not mobilise in the offspring. However, if the female does not have any P elements (M cytotype) and if the male has P elements, the elements will be able to transpose and cause hybrid dysgenesis. If the female has inactivated P elements (M' cytotype), the P elements does not transpose in the offspring. Cytotype regulation is thought to be due to a maternally inherited protein which prevents P elements from mobilising (Simmons *et al*, 2007). In the soma, Rio (1990) demonstrated that P element activity is regulated by preventing the removal of the last intron in the mRNA. The truncated protein serves as a repressor and regulates the activity of the element.

In this study, I sought to overcome this barrier to horizontal transmission. Attempts were made to germline transform the *yellow white* strain of *D. melanogaster* with the Juan and Pip1 elements. As these elements are members of a large

monophyletic group within the Jockey clade that appears to have its origin within mosquitoes, the first question addressed is simply will mosquito Jockey retroposons transpose in the *D. melanogaster* genome? The more interesting question was what would be the impact of an unregulated actively transposing element like Juan on an insect with a much more gene dense genome? Hybrid dysgenesis would be expected, but in contrast to P elements it would in theory persist in successive generations. The Pip1 element was more of an unknown and was included as a potential contrast to Juan.

Unaltered full length copies of Pip1, JuanA and JuanC were inserted into a piggyBac vector and together with a helper plasmid injected into *D. melanogaster* embryos. Lines containing JuanC and Pip1 were successfully established based on PCR detection of the elements in successive generations. The lines were each subdivided into five inbreeding populations and monitored for frequency of insects positive for the elements. As no selection was employed a progressive increase in frequency of positive insects was expected if the elements were active. The results were not all entirely consistent with expectations, so to resolve these difficulties the genomes of insects from different generations were sequenced.

5.2 Materials and Methods

5.2.1 Constructing the piggyBac Vector

The plasmid pXL-BacII was constructed by Li *et al* (2001). A 702bp fragment containing the terminal sequences of piggyBac was isolated by restriction enzyme digest. This fragment was ligated into pBlueScript II to form pXL-Bac II (Figure 5.1). The pGEM-T Easy Vector containing JuanA and JuanC were then digested with NotI (NEB) while the vector containing Pip1 was digested using EcoRI (NEB). pXL-BacII (Cary *et al*, 1989; Fraser *et al*, 1995) were also digested with the corresponding enzyme and ligated.

The plasmids were transformed into *E. coli* JM109 cells (Promega) and sent for resequencing to check for DNA integrity.

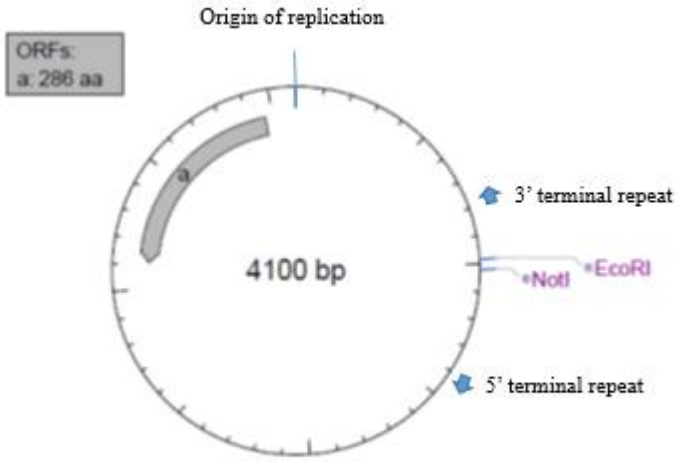

Figure 5.1. Representation of the plasmid pXL-BacII. The pXL-BacII plasmid is 4100bp in length. The piggyBac terminal repeats are at 800bp and 1400bp and the NotI and EcoRi sites are within the repeats. The single ORF is the ampicillin resistance coding domain.

5.2.2 Germline transformation

The following protocol was used. It is modified from Santamaria (1986) and Gompel (2005). *D. melanogaster* embryos from yellow white strains were used for germline injections. (1) Egg laying cages were set up the previous night and fresh embryos were harvested every 30 minutes from these cages. All subsequent steps were carried out at 18°C. The embryos were transferred, using a paintbrush, onto a microscope cover slip with double-sided sellotape and dechorionated by gently rolling the embryo on the tape; (2) Dechorionated embryos were aligned at the edge of the cover slip with the posterior pole pointing outwards. Leftover embryos were removed; (3) The cover slip was transferred into silica gel to dry for 4-6 minutes; (5) The cover slip was removed and the embryos were covered with a layer of halocarbon oil 700 (Sigma) and left for 5 minutes. (6) The cover slip with the embryos were mounted onto a microscope slide and positioned near the needle tip. Needles were

prepared using P-30 Vertical Micropipette Puller (Sutter) using 1.0mm OD borosilicate capillaries. (7) The piggyBac plasmid (1μg/μl) was coinjected with the helper plasmid pBSII-hs-orf (1μg/μl) (Cary *et al*, 1989; Fraser *et al*, 1995). The embryos were gently penetrated with the needle tip and the mix was injected into the embryos. (8)The needle was then quickly removed to reduce the amount of leakage. The process was repeated until all or most of the embryos were injected. (9) Excess halocarbon oil was drained from the cover slip. (10) The cover slip with the embryos were then transferred to a food vial and left at 25°C.

5.2.3 Establishment of *D. melanogaster* retroposon lines

The breeding design is summarised in Figure 5.2. Microinjected embryos were grown to adulthood and back-crossed to virgin yellow white flies. A single female was mated with 3 *yw* males while males were mated with 4 *yw* females. Virgin offspring (Generation 1, F_1) was collected before they were allowed to self-cross. Single females were isolated and placed in egg-laying tubes. Female flies were arbitrarily given alphabet and/or numerical names to enable lineage tracing. The offspring (F_2) of the cross was self-crossed again and single females were isolated and placed in egg-laying tubes. All the single females were PCR screened for the retroposon after they had laid sufficient number of eggs. Lines which were negative for two successive generations were discarded.

Each line was observed for any obvious phenotypic changes, such as eye colour and wing morphology. The egg hatch rate was also counted at generation 8. Females were allowed to lay eggs overnight in an egg laying dish and the number of eggs were counted. The number of unhatched eggs was scored after 24, 48 and 72 hours.

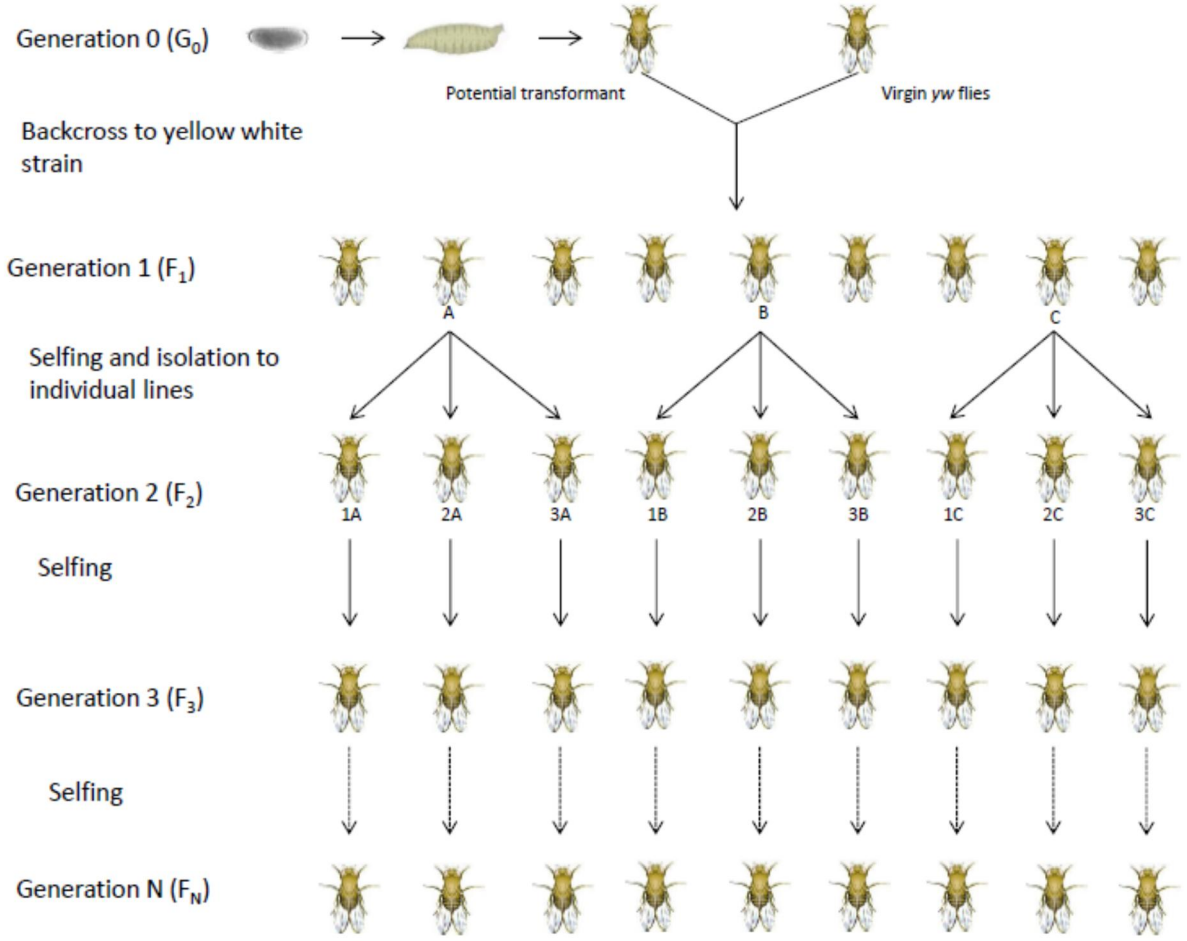

Figure 5.2 Diagram of the establishment of transformed fly lines. Full description is in section 5.2.3. The microinjected embryos were grown to adulthood and backcrossed to virgin yellow white flies. Generation 1 of the cross was selfed and the single females were isolated and placed in egg-laying tubes. The offspring (G_2) of the cross was selfed again and single females were isolated and placed in egg-laying tubes. Female flies were arbitrarily given alphabet and/or numerical names to enable lineage tracing.

5.2.4 Next generation whole genomic sequencing

10 adult transformed *Drosophila melanogaster* flies were sent for whole genomic MiSeq DNA sequencing. The sequencing was performed by the Genome Centre (Charterhouse Square, Queen Mary). The flies were selected from those lines that appeared to be fixed for an element. 3 were from the JuanC line 4F while the other 7 were from Pip1 line 2E. 3 of the Pip1 lines presented themselves with the dark pigmented eyes. The flies were taken from Generation 3, 8 and 14. This design would

allow the movement (if any) of the retroposon to be tracked. The coverage was 7x (Table 5.1).

Table 5.1 Flies sent for MiSeq DNA sequencing. 3 were from the JuanC lines while the rest were from Pip1 lines. Flies from different generations were chosen to obtain a better view of mobile element movement.

Fly	Insertion	Line	Generation	Coverage
1	JuanC	4F	3	7x
2	JuanC	4F	8	7x
3	JuanC	4F	14	7x
4	Pip1	2E	3	7x
5	Pip1	2E	8	7x
6	Pip1	2E	14	7x
7	Pip1	2E	14	7x
8	Pip1	2E (dark pigmented eyes)	3	7x
9	Pip1	2E (dark pigmented eyes)	8	7x
10	Pip1	2E (dark pigmented eyes)	14	7x

Two different approaches were used to analyse the NGS read data. The first was to generate contigs using a genomics workbench tool, CLC Genomics Workbench 7.5 (http://www.clcbio.com/products/clc-genomics-workbench/). A summary of the contigs generated is presented in Figure 5.3. The contigs were then screened using BLAST (implemented in the Workbench) to identify contigs with hits to the retroposon sequence.

1.3 Contig measurements (excluding scaffolded regions)

N75	403
N50	489
N25	604
Minimum	143
Maximum	10,004
Average	470
Count	316,658
Total	148,800,899

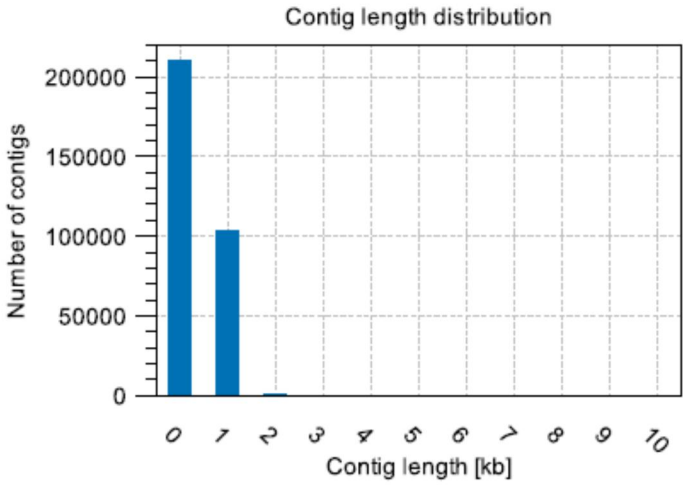

Figure 5.3 A summary of the contigs generated. Most of the contigs generated were less than 1kb.

The second approach was analysing the reads without generating contigs, in case informative reads were not being included in the assembly. Data analysis was performed on the Galaxy platform (Giardine *et al*, 2005; Blankenberg *et al*, 2010; Goecks *et al*, 2010). The raw reads from the sequencing was pre-processed for quality (Figure 5.4A). The steps were: (1) The read file was converted into a usable format. FASTQ groomer was used to convert the FASTQ files into *sanger* format files (Blankenberg *et al*, 2010). (2) The reads were clipped to remove adapter sequences (http://hannonlab.cshl.edu/fastx_toolkit/galaxy.html). (3) A FASTQC report file was generated for initial checking of the read quality (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). (4) Reads that did not meet the minimum quality were modified or removed by trimming (http://hannonlab.cshl.edu/fastx_toolkit/galaxy.html). (5) Another FASTQC report file was generated to check the quality of the reads. If the reads were still unsatisfactory,

steps (2)-(5) was repeated. The amount of read output was different for the forward and reverse primer sequencing files after the quality checks. In order to check for any associated bias, the two types of files were used separately.

After pre-processing, the reads were aligned to the *Drosophila melanogaster* genome (build 3) using Bowtie2 (Langmead *et al*, 2012). These generated two files: Reads which aligned and reads which did not align to the genome. Both files were then aligned to the retroposon sequence and again, it generated files which contain aligned and unaligned files. A summary of this workflow is in Figure 5.4B.

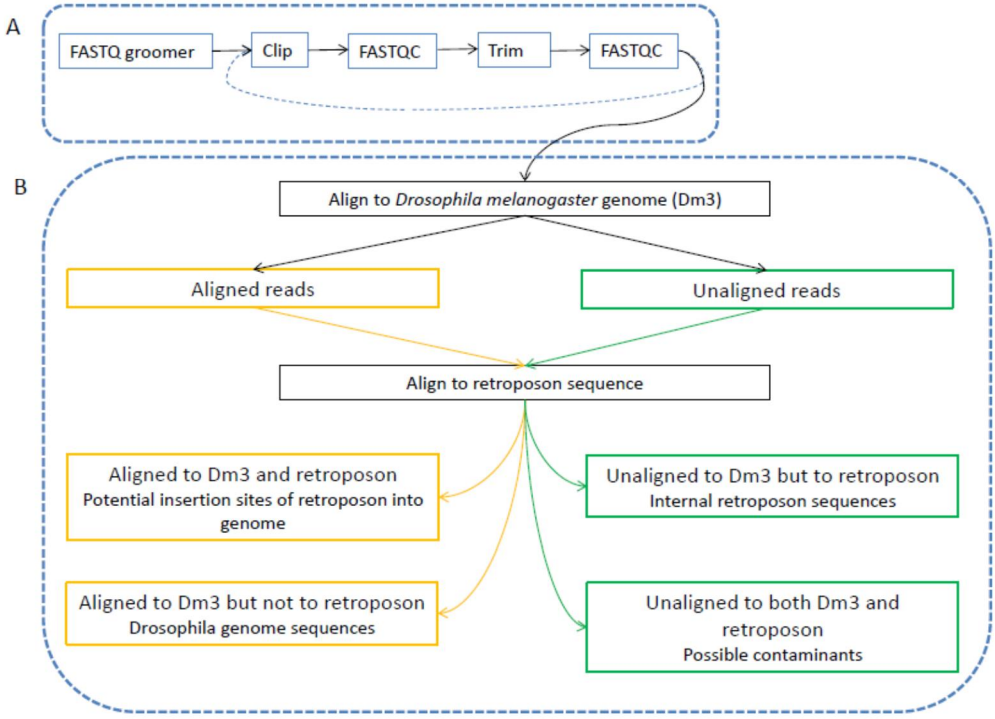

Figure 5.4 Workflow of the analysis performed using Galaxy. A) Workflow of the pre-processing quality checks done on the reads. The files were converted to FASTQ files first before they were clipped and trimmed. The FASTQC step checks the quality of the reads. B) Workflow of the alignment. The reads were first aligned to the *Drosophila melanogaster* genome (build 3) using Bowtie2. This generates an aligned and an unaligned file. Both files were then separately aligned to the retroposon sequence and again, this generates an aligned and an unaligned file. The potential identities of the reads are listed in the diagram.

5.3 Results

5.3.1 Successful hatchlings

Table 5.2 summarises the number of injected embryos, successful hatchlings and the adults obtained from the germline injections.

Table 5.2 Summary of the number of embryos infected, larvae and adults. Percentage in brackets represents the hatch rate and survival to adult rate respectively.

	Pip1	JuanA	JuanC
Number of injected embryos	325	493	378
Hatchlings	26(8.0%)	37 (7.5%)	34(10.4%)
Adulthood	13 (50%)	16(43.2%)	12(35.3%)

5.3.2 Verification of transformation

From PCR results, an adult containing Pip1 and JuanC was successfully obtained from the transformation, while JuanA injections did not produce a transformed fly (data not shown). Offspring from the transformed adults were self-crossed and the number of positive individuals is presented in the Figure 5.5-5.7.

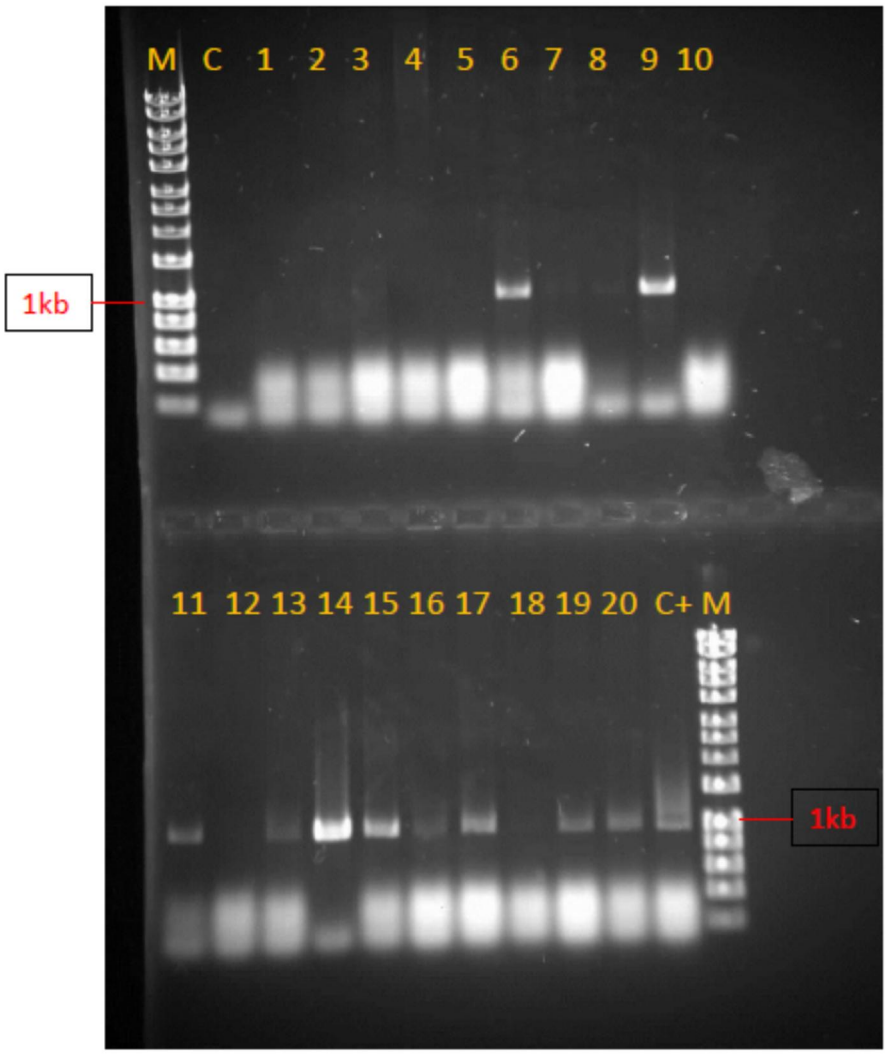

Figure 5.5 Gel electrophoresis result of the PCR screening for Pip1 from 10 individuals per generation. The PCR primers amplify 1kb. The flies are from generation 5. Flies 1-10 are from line 5A while flies 11-20 are from 2E. M is the Bioline HyperLadder 1 (Bioline, UK). C is a negative control for DNA while C+ is a positive control. The 1.0% Agarose gel was ran at 90V for 2 hours.

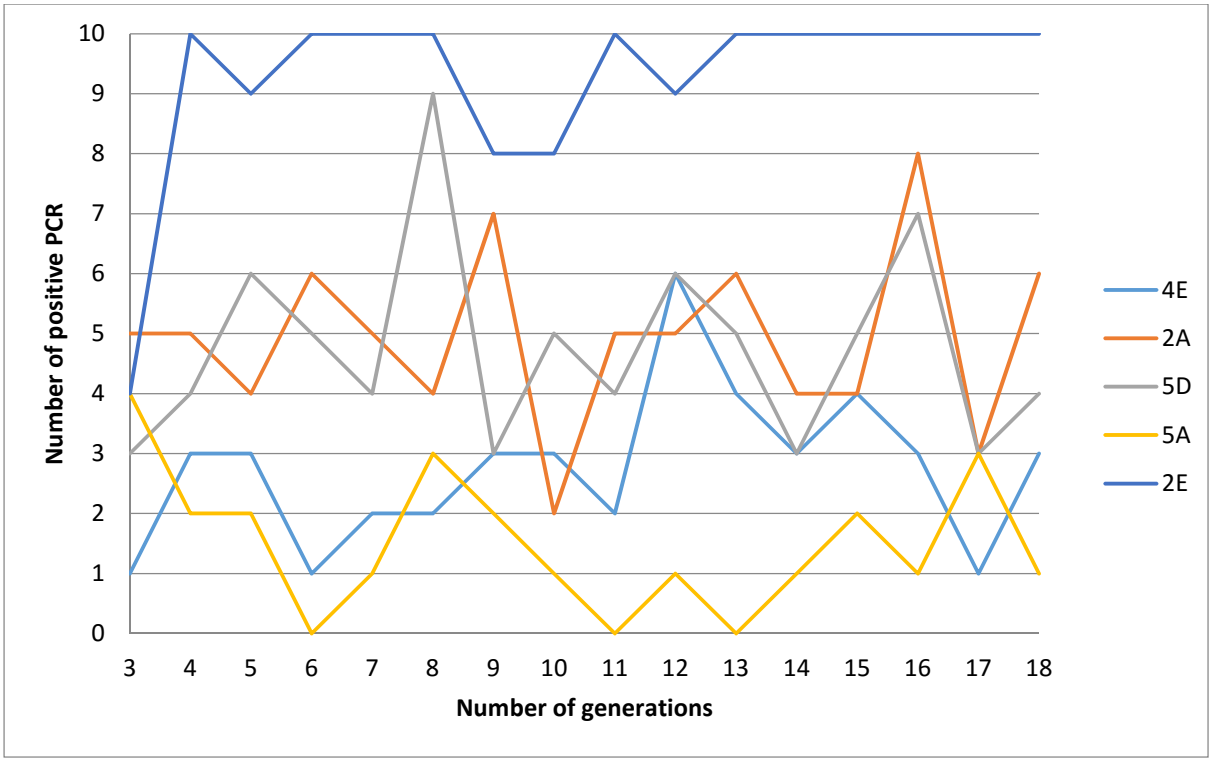

Figure 5.6 Number of positive Pip1 individuals per generation in different fly lines established. Line codes are given in the key

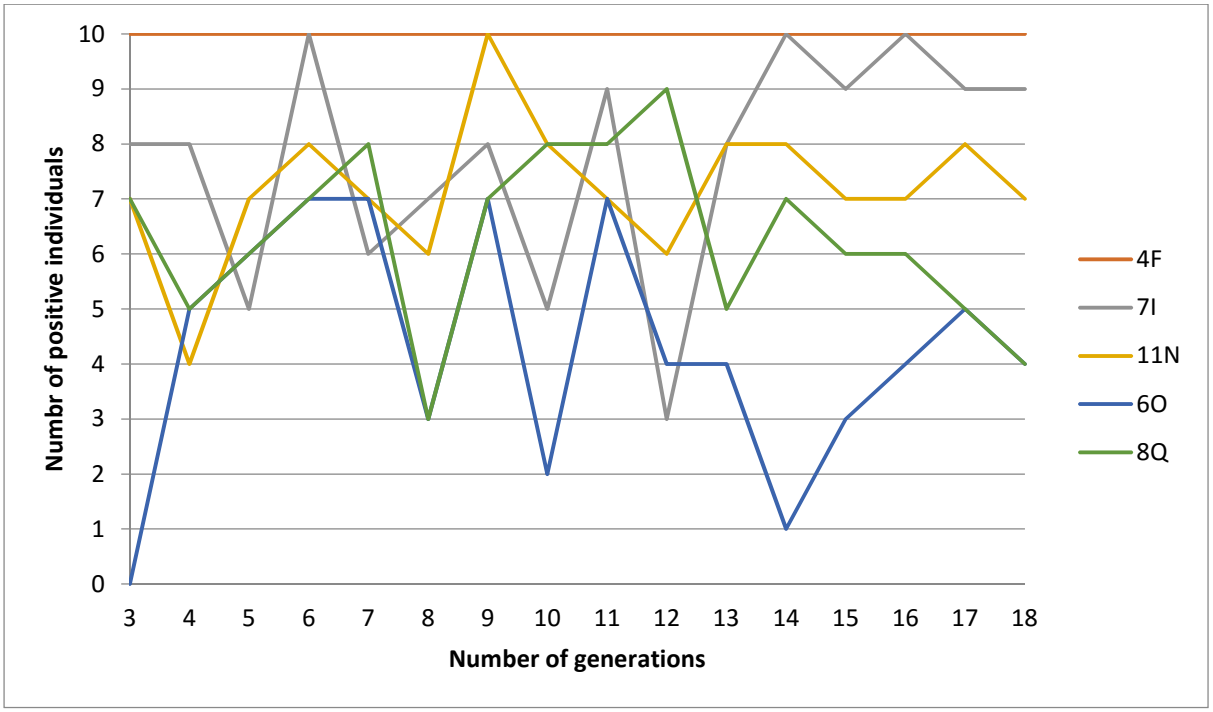

Figure 5.7 Number of positive JuanC individuals per generation in different fly lines established. Line codes are given in the key

5.3.3 Phenotypic mutations observed

A common effect of transposable element mobilisation is phenotypic changes to the flies. I sought to characterise phenotypic changes by changes to egg hatch rates as well as general observation of the flies. The egg hatch rate for Pip1 and JuanC fly lines are presented in Table 5.3 and 5.4 respectively. The baseline hatch rate for the yellow white flies was about half (50.64%). For the Pip1 lines, two of the matings went below this figure (2E *x* 2E and 5D *x* yw). The hatch rates for the other matings and all the JuanC lines were higher than this.

Table 5.3 Hatch rate of Pip1 fly lines. Virgin females were crossed with virgin males.

Female	Male	Total eggs laid	Total hatched eggs	Hatch rate (%)
yw	yw	472	239	50.64
2E	2E	703	322	45.80
2E	yw	493	300	60.85
yw	2E	1048	632	60.31
5A	5A	506	271	53.56
5A	yw	638	390	61.13
yw	5A	493	355	72.01
2A	2A	966	535	55.38
2A	yw	446	298	66.82
yw	2A	600	291	48.50
4E	4E	595	452	75.97
4E	yw	421	303	71.97
yw	4E	588	413	70.24
5D	5D	478	273	57.11
5D	yw	600	291	48.50
yw	5D	525	342	65.14

Table 5.4 Hatch rate of JuanC fly lines. Virgin females were crossed with virgin males.

Female	Male	Total eggs laid	Total hatched eggs	Hatch rate (%)
4F	4F	803	606	75.47
4F	yw	865	677	78.27
yw	4F	706	438	62.04
8Q	8Q	822	447	54.38
8Q	yw	618	461	74.60
yw	8Q	728	528	72.53
7I	7I	530	419	79.06
7I	yw	641	483	75.35
yw	7I	538	431	80.11
11N	11N	529	430	81.29
11N	yw	448	343	76.56
yw	11N	630	445	70.63
6O	6O	509	371	72.89
6O	yw	616	421	68.34
yw	6O	569	387	68.01

5.3.4 Mosaicism in fly eye pigmentation

While there was no significant changes to egg hatch rate, the Pip1 2E line presented adult flies with dark randomly pigmented eyes (Fig 5.8 and 5.9). The distribution of pigments showed no clear pattern, was different from the left to right eye of any one individual, and also differed from individual to individual.

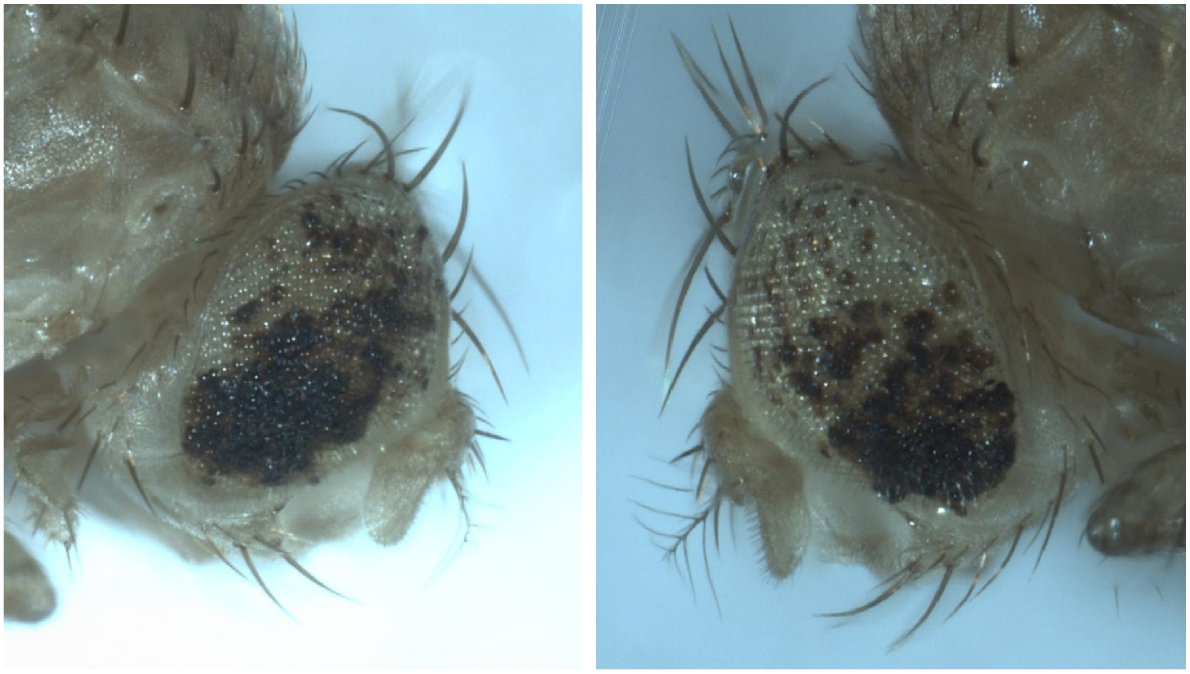

Figure 5.8 Dark eye pigmentation observed in a Pip1 transformed male *D. melanogaster*. Lateral view of the eye from a single male. Note the different spot profile between the right and left eye.

Figure 5.9 Light eye pigmentation observed in a Pip1 transformed male *D. melanogaster*. Lateral view of the eye from a single male. Note the different spot profile between the right and left eye, and the different intensity of the dark spots between this individual and the previous male fly (Fig 5.8).

The mosaic flies was first observed in the third generation and the mosaic individuals were collected and isolated for 6 generations (Table 5.5). The total number of mosaic individuals collected increased in the 5th generation compared to the previous generation. This is because more adults were kept and allowed to breed for the next generation and allowance of longer emerging time for the pupae. Fly vials were normally discarded 2 weeks after the first adult emerged but an extra week was added in order to increase screen more adult flies for mosaic individuals.

Table 5.5. The number of mosaic individuals collected per generation.

Generation	Males	Females	Total	Total adults screened	Percentage of mosaic flies from (%)
3	23	31	54	594	11
4	15	26	41	410	10
5	40	43	83	1162	14
6	35	46	81	1053	13
7	39	40	79	869	11
8	31	37	68	816	12

To determine the heritability of this trait, mosaic flies were isolated and crossed with each other as well as to the parental yellow white strain. Virgin flies were collected and mosaic males were mated to mosaic females in a 3:1 ratio. Mosaic males was back crossed with yellow white females in a 1:4 ratio while yellow while males were crossed with mosaic females in a 1:3 ratio. However, mosaic individuals suffer from a reduced life span and did not mate readily. Some of the virgin mosaic flies collected died before the crosses could be set up. Most of the crosses set up did not

produce any offspring. Offspring from successful crosses were kept and observed for the mosaic phenotype but this phenotype was not observed neither in the first nor the second generation of the crosses (data not shown).

Analysis was carried out on the whole genome sequencing of three individuals with mosaic eye pigmentation. A more in-depth analysis of the genome sequencing data is described in the next section (5.3.4), so the analysis described here focuses on identifying a possible genetic basis for the phenotype observed. Sequencing reads that contain both D. melanogaster sequences and Pip1 sequences were analysed to determine where Pip1 might have inserted. However, the *D. melanogaster* sequences in these reads were too short to enable identification of the insertion sites.

5.3.5 NGS data analysis

The BLAST screen of the contigs generated produced numerous hits. However, only most of the reads which aligned to Pip1 were very short. Only one of the hits produced a significant hit to Pip1. An example of the BLAST run is in Figure 5.10.

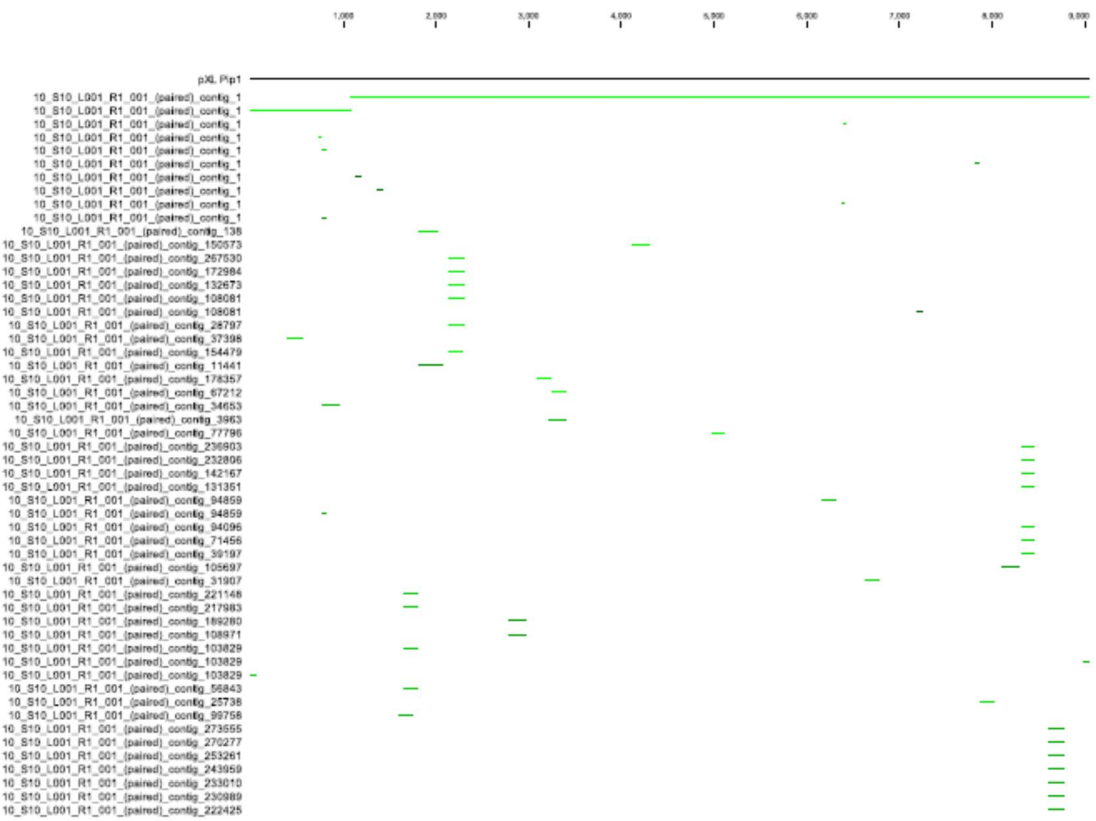

Figure 5.10 BLAST output for the contigs against Pip1. The output was generated from CLC Genomics Workbench.

The results of the Galaxy analysis are presented in Table 5.6. The overall alignment to Dm3 and to the retroposon is presented. Most of the reads sequenced are *D. melanogaster* sequences; however, about two-thirds to half of the reads do not align to either *D. melanogaster* or to the insertion vector. The reads that aligned to both the *D. melanogaster* genome and retroposon sequences were analysed. However, the putative *D. melanogaster* sequences in these reads too short to have a single exact location in the genome.

On the other hand, the sequences with matches to the retroposon included reads that also clearly matched the piggyBac vector. It was possible to assemble a full plasmid construct from these reads. This result raises two possibilities to explain the flies in which the PCR assay detected the retroposon: the whole plasmid could have been integrated into the fly chromosome or the plasmids could be present as extrachromosomal DNA.

A histogram of read-depth against plasmid nucleotide position was constructed for each fly. An example is given in Figure 5.11. The read frequencies are unequally distributed along the whole plasmid. Pip1 starts around 1kb and ends around 5.5kbp.

Table 5.6 Analysis of NGS reads. The table shows the overall alignment to *Drosphila melanogaster* build 3, the overall alignment to the retroposon (either JuanC or Pip1) and the percentage of sequences that does not align to both.

Fly	Insertion	Overall alignment to Dm3 (%)	Overall alignment to retroposon (%)	Does not align (%)
1	JuanC	63.09	0.94	36.59
2	JuanC	63.79	0.95	35.90
3	JuanC	51.20	1.28	48.36
4	Pip1	57.09	2.07	41.47
5	Pip1	60.91	2.08	37.67
6	Pip1	60.91	1.79	37.86
7	Pip1	64.62	2.03	34.00
8	Pip1	46.47	2.86	51.58
9	Pip1	70.28	1.68	28.59
10	Pip1	48.76	2.63	49.41

Figure 5.11 Histogram of reads against plasmid nucleotide position in Fly 4 Forward. Pip1 is located from 1kbp onwards to 5.5kbp. The minimum read value was 150 reads while the maximum was 567.

5.4 Discussion

The retroposons JuanC and Pip1 was introduced into *Drosophila melanogaster* yellow white strains. Despite numerous attempts to obtain more transformants (by increasing the number of injected embryos) only one successful transformant was obtained for each of Pip1 and JuanC while we did not manage to obtain a JuanA transformant.

The lines were inbred to obtain a homozygous population. The frequency of the retroposon changed in the different lines, and in one of the lines, it seems to have been fixed. I did not actively select for the retroposon but allowed the individuals to mate at random. However, each line was established from a single F1 and F2 female, thus producing a bottleneck/founder effect to increase the chances of obtaining a homozygous population.

I surveyed the lines for evidence of hybrid dysgenesis. This term refers to the high rate of mutations observed in the offspring of crosses between two strains. Early reports of this phenomenon were from crosses between *Drosophila melanogaster* strains: and subsequent research showed that it was due to the mobile 'P elements'; when a P cytotype male was mated with a naïve or M cytotype female dysgenesis occurred, including sterility and death (Engels, 1989).

While I did not observe evidence of reduction in egg hatch rates, a phenotypic mutation developed in the flies after a few generations. Individuals, both male and female, have black spots on their eyes and the degree of spots varies between the left and right eye and among individuals, ranging from mildly spotted (Fig 5.9) to roughly 50% spotted (Fig 5.8). Efforts to establish fly lines were unsuccessful despite numerous attempts to breed the flies. About a third of individuals with 50% or more black spots did not survive more than 2 days after eclosion. Flies that survive do not breed readily and it was difficult to obtain offspring. From the flies which bred, this phenotype was neither present in the offspring nor the subsequent generation.

The different pattern observed within and between individuals and the lack of heritability of the trait suggests that the mutation was present at the eye cells rather than the whole fly. The *Drosophila melanogaster* strain used in this study carries a P element insertion in the eye colour gene, hence it is unable to produce the red colour pigmentation in the eye and displays a white coloured eye phenotype. A spontaneous reversion event where the P element mobilises itself from the eye colour gene was ruled out as unlikely (Prof. Stanewsky, pers. comm.).

A possible event is the Pip1 element might have cross mobilised the P element from the white eye gene and caused a partial reversion to the red eye phenotype. Transposable elements have been known to cross-mobilise other elements in the genome in the offspring of a hybrid cross. Petrov *et al* (1995) found that different classes of mobile elements were mobilised in *D. virilis* in the offspring of a dysgenic cross. The authors found that cross-mobilisation is possible if an element complements the functions of another element, thus allowing cross-mobilisation to take place. In this study, the functional Pip1 element might complement the function of the transposase of the P element and cause it to mobilise from the gene and produce a partial reversion to red colour phenotype.

The DNA quality was subjected to rigorous testing to ensure it was free of contamination. The DNA extraction protocol was optimised to maintain the integrity of the DNA as well as reducing the chances of contamination. The amount of DNA was measured and quality testing was performed and only DNA which met the quality standard was used in the sequencing reactions. Both positive (using diluted plasmid DNA with the retroposon sequence) and negative controls (no DNA or plasmid DNA without the retroposon sequence) were used during the PCR to validate interpretation of the results and were consistent. For example in Figure 5.5, there were no bands obtained in the negative control but a band is present in the positive control. In addition, the DNA from all the samples were extracted at the same time. The results showed that some of the lanes produced the expected band while others did not.

Therefore, it is unlikely that foreign DNA was introduced during DNA extraction or running the PCR. The results obtained from determining the number of positive individuals per generation also suggests that DNA contamination is unlikely (Fig 5.6 and 5.7). DNA from 10 individual flies from different breeding lines were extracted at the same time. Some of the individuals within a line were positive but some were not (except for line 4F for JuanC, which consistently showed positive results for all 10 individuals).

The NGS data confirmed the presence of the retroposon in the fly lines. While I did not detect any junction of the retroposon to fly DNA, this result does not eliminate the possibility that a junction is present but our NGS strategy did not pick it up. The 7x coverage might not be enough to sequence the whole genome sufficiently. As with most genomic sequencing, the coverage was not even over the genome, and telomeres and repetitive regions are hard to sequence and assemble. Hence insertions into these regions might have remained undetected especially given the shortened read-length after quality control.

The piggyBac-retroposon plasmid could possibly be present as extrachromosomal DNA in the cell. This would explain even read depth observed over the whole length of the plasmid/retroposon contrstuct, in contrast with the absence of contigs extending into *D. melanogaster* DNA. It is not uncommon for injected plasmids to be present in the organism *D. melanogaster* (Spradling and Rubin, 1982). In an experiment with P elements, Spradling and Rubin found out that the P elements transposed from extrachromosomal injected plasmids into the *D. melanogaster* chromosome. In addition, extrachromosomal circular DNA are found in various organisms, including *D. melanogaster* (Cohen *et al*, 2009) and consist mainly of tandemly repeated genomic sequences.

If the plasmid had become established as an extrachromosomal element, then it could be inherited vertically from mother to offspring. Random fluctuations in the

contribution of each matriline would lead to a drift in the frequency of the element, which is consistent with the fluctuations in population frequency observed in the fly lines. *D. melanogaster* contains autonomously replicating sequences (Marunouchi and Hosoya, 1984). These sequences are capable of initiating replication at replication origins independent of cell control, hence the name. Brun *et al* (1990) found that a stretch of 800kbp on the *D. melanogaster* X chromosome is capable of promoting autonomous replicating ability in *Saccharomyces cerevisiae*. Therefore, the plasmid might be present as extrachromosomal element in the fruitflies.

The plasmid might also have been taken up by the bacterial endosymbiont *Wolbachia*. *Wolbachia* have been identified in most insect species, including *D. melanogaster*, and have been implicated in causing disease resistance and host reproduction (Hurst *et al*, 1999; Hedges *et al*, 2008). The plasmid could have been taken up by *Wolbachia* and replicate within the bacteria. In addition, the origin of replication of *Wolbachia* and *E. coli* is similar (Hotopp et al, 2007). Thus, it is possible that *Wolbachia* could replicate the plasmid.

It remains possible that transposition has occurred in the somatic tissue, explaining the eye pigmentation phenotype associated with low fitness. Future work could follow up this possibility by obtaining sequence from the eye tissue of flies exhibiting this phenotype, and following the protocol set out above to assay the PCR-positive lines. If there had been transposition, that could be detected by the occurrence of sequences integrated into the fly genome.

CHAPTER 6

CONCLUDING REMARKS

This thesis has explored retroposons and their potential applications in mosquito genomics research. Our knowledge of how retroposons behave and interact with their host genome is still patchy. Research has mainly focused on Class II elements, the transposons. However, retroposons – such as the human LINE-1 and the Juan elements targeted in this study – are found in high copy numbers and deserve comparable attention. Similarly, research on the genomic composition of fruitflies has progressed in leaps and bounds, but mosquito genomic research is still lagging behind. This discrepancy should be rectified, especially considering that mosquitoes are vectors of many deadly diseases.

As an initial step to develop the tools to probe the genomics of mosquitoes, I have identified characterised the Juan elements and Pip1, highlighting their copy number differences, similarity in conserved motifs at their coding domains and tested their usefulness in germline transformation by injecting them into *Drosophila melanogaster*.

The main experimental findings of the thesis were summarised within their respective chapters: I have investigated retroposons and the mosquito genome; characterisated Pip1 and developed artificial horizontal transfer of retroposons. However, a few main themes became apparent in the course of the thesis.

Firstly, the sequenced mosquito genomes are quite distinctive. The anophelines have maintained a low TE content in comparison to its genic content (Holt et al, 2007; Marinoti et al, 2013). On the other hand, the *Aedes aegypti* genome has ballooned in size and almost half of it are TEs (Nene et al, 2007). No other sequenced insect groups

show this remarkable diversity in genome sizes. In addition, the organization of the genome has changed from a pattern of long stretches of unique genes interrupted by non-coding DNA to the complete opposite pattern- long stretches of non-coding DNA with genic sequences in between. Therefore, this group makes an excellent case study to understand evolution of genome organization. Furthermore, the high copy number retroposons, the Juan elements, are only present in the *Culex* and *Aedes* genome. Combining this information with other research (Bohne et al, 2007; Belyayev 2014), it is possible that the burst of activity by the Juan elements could have driven speciation of *Culex* and *Aedes* mosquitoes. A burst of retroposon activity would have caused major restructuring in genome organization, as evidenced by the change in genome interspersion pattern.

Secondly, in contrast to high copy-number transposable elements, there are low copy number elements, such as Pip1. My work found that Pip1 insertion sites are polymorphic between strains, verifying the initial reports of Crainey and Malcolm (2007). This observation suggests that Pip1 has been active in recent evolutionary time. Pip1 likely arose prior to the geographic spread of the *Culex* genus and has continued to be active in the different *Culex* strains. It also fits the master gene hypothesis to some extent. Its transcription generates multiple copies, some which display truncations near the same region due to formation of secondary structures.

It would increase our knowledge of retroposon biology if Pip1 activity could be further tracked either in the transformed *Drosophila melanogaster* or in the *Culex quinquefasciatus* Johannesburg strain. If active copy number decreases, that would suggest that it is difficult for a retroposon to survive, even in a large genome offering plenty of safe insertion sites. This type of study would provide insights on how a retroposon is deactivated and controlled in a genome. Alternatively, if Pip1 is able to increase in copy number and reach a comparable frequency to JuanC, this would provide insights into how the genome tolerates high copy number elements and

whether there is a threshold point above which an element can increase in copy number dramatically (i.e. a tipping point).

Germline transformation is an exciting tool in biology. It provides the means to introduce a gene from one species to another. The introduction of Pip1 into *D. melanogaster* was achieved using this technique (although it appears not to have established in the germ line). The transformed *D. melanogaster* strains showed a change in eye colour phenotype but not in egg hatch rate. By comparison, when P elements invaded *D. melanogaster*, a variety of phenotypes was observed, including reduced fecundity and random mutations (Engels, 1989). These differences suggest that effects of transposable elements depend critically on the nature of the element.

There are now a large number of databases storing whole genomic DNA sequencing data of recently sequenced species. In addition, many more species are in the pipeline to have their genome sequenced. Throughout my PhD, I have used Repbase (database for transposable elements), Genbank (a collection of publicly available DNA sequences), and VectorBase (database for medically important pathogen-carrying organisms). However, an area that still requires improvement is validating the assembled gene sequences. On a number of occasions the sequences encountered in these databases were incorrect. Research programmes to continually validate and revise revise gene assembly should be implemented to check the entries into databases to and reduce this type of confusion.

The focus of this research has been on the Juan elements and Pip1 in the mosquito genome. There are plenty of other retroposons waiting to be explored and characterised in the genome. These elements were chosen based on their specific unique characteristics- namely they are retroposons making up a major genomic component of mosquito genomes, they are elements that have recently been active in their host, and they provide a contrast of the behaviour of a high versus low copy number element.

There is still plenty to discover and investigate about retroposons and mosquito genome. Retroposon biology still remains understudied, with most extisting work having addressed human LINE-1 elements. Mosquito genomics are also understudied, with the focus being mainly on generating transformed strains which reduce disease transmission. Little research is done to understand the various genes in the mosquitoes, let alone transposable elements which contribute so much to the evolution of the genome.

Looking back at the course of the PhD with the advantage of hindsight, I would have taken different approaches at a few junctions. Firstly, I would have attempted to synthesize the retroposons artificially. A lot of the initial laboratory work involved molecular cloning of the retroposons. Due to the polymorphic nature of the elements, I had to wait for new mosquito strains to arrive before suitable genomic DNA could be obtained. I did explore this avenue, but the cost to synthesize a single element was in the price range of £3000, which is a substantial investment, and thus, this approach was not pursued.

Secondly, an attempt to clone the retroposon into a piggyBac vector containing fluorescent markers, and also inserting the fluorescent marker into the piggyBac vector containing the retroposons was made. However, the *E. coli* colonies grown did not contain a transformed plasmid despite numerous attempts. This is likely because including the marker increased the size of the plasmid from 7kb to 11kb. It appears that the larger size was sufficient to have reduce the viability of the bacteria and hence, the bacteria with this larger plasmid were not obtained. Due to these constraints and insufficient time, the transformation was carried out without a genetic marker. The PCR screening was effective, but with a fluorescent marker, screening of fruitflies would have progressed much faster.

Additional experiments I would carry out would be the germline transformation on the mosquitoes, especially the Juan elements on *Anopheles*

mosquitoes. Since both *Aedes* and *Culex* mosquitoes possess the Juan elements, it would be of high research value to find out if the Juan elements can achieve a high copy number in the *Anopheles* genome.

Expression of retroposon proteins was also attempted in order to emulate the work of Eickbush *et al* (2000). Some success was achieved producing Pip1 proteins from *E. coli* expression cells, but the Juan proteins were difficult to express. Binding the protein to mRNA also was difficult despite using different incubation protocols and collaboration another group with experience in such studies.

I have been able to take the first steps in studying retroposons in mosquito genomes: certain retroposons are present in unusually high abundance; and it is possible to introduce a retroposon into another species using artificial means. Additional exploration of retroposon proteins on top of the information gleaned from genomic studies would enhance the field further.

REFERENCES

Adams M.D. *et al*, 2000, The Genome Sequence of *Drosophila melanogaster*, *Science*, 287, 2185-2195

Agarwal M., Bensaadi N., Salvado J.C., Campbell K., Mouchès C., 1993, Characterization and Genetic Organisation of Full-Length Copies of a LINE Retroposon Family Dispersed in the Genome of *Culex pipiens* Mosquitoes, *Insect Biochem. and Molec. Biol.*, 23, 621-629

Arensburger *et al*, 2010, Sequencing of *Culex quinquefasciatus* Establishes a Platform for Mosquito Comparative Genomics, *Science*, 330, 86-88

Beck C.R., Collier P., Macfarlane C., Maliq M., Kidd J.M., Eichler E.E., Badge R.M., Moran J.V., 2010, LINE-1 Retrotransposition Activity in Human Genomes, *Cell*, 141, 1159-1170

Beck C.R., Garcia-Perez J.L., Badge R.M., Moran J.V., 2011, LINE-1 Elements in Structural Variation and Disease, *Ann. Rev. Genomics Human Gen.*, 12, 187-215

Beisel C., Imhof A., Greene J., Kremmer E., Sauer F., 2002, Histone Methylation by the *Drosophila* Epigenetic Transcriptional Regulator Ash1, *Nature*, 419, 857-862

Belyayev A., 2014, Bursts of Transposable Elements as an Evolutionary Driving Force, J. of Evo. Biol., doi 10.1111

Bensaadi-Merchermek N., Cagnon C., Desmons I., Salvado J.C., Karama S., D'Amico F., Mouchès C., Agarwal M., 1997, CM-gag, a Transposable-like Element Reiterated in the Genome of Culex pipiens Mosquitoes, Contains only a *gag* Gene, Genetica, 100, 141-148

Betram G., Innes S., Minella O., Richardson J.P., Stansfield I., 2001, Endless Possibilities: Translation Termination and Stop Codon Recognition, *Microbiology*, 147, 255-269

Biedler J.K. and Tu Z., 2007, The Juan non-LTR Retrotranposon in Mosquitoes: Genomic Impact, Vertical Transmission and Indications of Recent and Widespread Activity, *BMC Evolutionary Biol.*, 7, number 112

Blair C.D., Adelman Z.N., Olson K.E., 2000, Molecular Strategies for Interrupting Arthropod-Borne Virus Transmission by Mosquitoes, *Clinical Microbiology Rev.*, 13, 651-661

Blankenberg D., Gordon A., Von Kuster G., Coraor N., Taylor J., Nekrutenko A., Galaxy Team, 2010, Manipulation of FASTQ data with Galaxy. *Bioinfo.,* 26, 1783-1785.

Blankenberg D., Von Kuster G., Coraor N., Ananda G., Lazarus R., Mangan M., Nekrutenko N., Taylor J., 2010, Galaxy: A Web-based Genome Analysis Tool for Experimentalists, 19: Unit 19.10.1-21.

Bohne A., Brunet F., Galiana-Arnoux D., Schultheis C., Volff J., 2008, Transposable Elements as Drivers of Genomic and Biological Diversity in Vertebrates, Chromosome Research, 16, 203-215

Boissinot S., Chevret P., Furano A.V., 2000, L1 (LINE-1) Retrotransposition and Amplification in Recent Human History, *Molec. Biol. and Evo.*, 17, 915-928

Brun C., Dang Q., Miassod R., 1990, Studies of an 800-kilobase DNA Stretch of the Drosophila X Chromosome: Compaaing of a Subclass of Scaffold-Attached Regions with Sequences Able to Replicate Autonomously in Saccharomyces cerevisiae, *Mol. Cell. Biol.*, 10, 5455-5463

Cary, L.C., Goebel, M., Corsaro, B.G., Wang, H.G., Rosen, E., Fraser, M.J., 1989. Transposon Mutagenesis of Baculoviruses: Analysis of *Trichoplusia ni* Transposon IFP2 Insertions within the FP-locus of Nuclear Polyhedrosis Viruses, *Virology*, 172, 156-69.

Catteruccia F., Nolan T., Loukeris T.G., Blass C., Savakis C., Kafatos F.C., Crisanti A., 2000, Stable Germline Transformation of the Malaria Mosquito *Anopheles stephensi*, *Nat.*, 405, 959-962

Charleswoth B., Langley C.H., 1989, The Population Genetics of *Drosophila* Transposable Elements, *Annu. Rev. Genet.*, 23, 251-287

Cohen S., Yacobi K., Segal D., 2003, Extrachromosomal Circular DNA of Tandemly Repeated Genomic Sequences in *Drosophila*, Genome Research, 13, 1133-1145

Crainey J.L and Malcolm C.A., 2010, Retrotransposon Insertion Sites Vary within and between Populations of *Culex pipiens* form molestus, *Annals of Tropical Med. And Parasitology*, 104, 355-358

Crainey J.L., Garvey C.F., Malcolm C.A., 2005, The Origin and Evolution of Mosquito APE Retroposons, *Molec. Biol. Evo.*, 22, 2190-2197

Darboux I., Charles J., Pauchet Y., Warot S., Pauron D., 2007, Transposon-mediated Resistance to *Bacillus sphaericus* in a Field-evolved Population of *Culex pipiens* (Diptera: *Culicidae*), *Cellular Microbiology*, 9, 2022-2029

Dawkins R., 1976, *The Selfish Gene*, Oxford University Press

de Koning A.P., Gu W., Castoe T.A., Batzer M.A., Pollock D.D., 2011, Repetitive Elements May Comprise Over Two-Thirds of the Human Genome, *PLoS Gen.*, 12, e1002384

Dragan M., Moghul M.I., Priyam A., Wurm Y., 2014, GeneValidator: Identify Problematic Gene Predictions, *in prep*

Durand P.M., Oelofse A.J., Coetzer T.L., 2006, An Analysis of Mobile Genetic Elements in Three *Plasmodium* Species and Their Potential Impact on the Nucleotide Composition of the *P. falciparum* Genome, *BMC Genomics*, 7, 282

Edi C.V.A., Kondou B.G., Jones C.M., Weetman D. and Ranson H., 2012, Multiple-Insecticide Resistance in *Anopheles gambiae* Mosquitoes, Southern Cote d'Ivoire, *Emer. Infect. Dis.*, 18, 1508-1511

Eickbush T.H., 2002, R2 and Related Site-specific Non-long Terminal Repeat Retrotransposons, in *Mobile DNA II*, edited by Craig N.L., ASM Press, Washington, D.C., 813-35

Eickbush, T., and Malik, H., 2002. Origins and Evolution of Retrotransposons, in *Mobile DNA II*, eds Craig N., Craigie R., Gellert M., Lambowitz A., American Society for Microbiology, Washington, D.C., 1111-44

Engels, W. R., 1989, P elements in *Drosophila melanogaster*, Berg D.E. and Howe M.M. (eds.) *Mobile DNA*, American Society for Microbiology, 437-484

Engels, W.R., 1989. P elements in *Drosophila*, in *Mobile DNA*, eds. Berg, D.E., and Howe, M.M., American Society for Microbiology, Washington D.C, 437-84

Fraser M.J., Cary L., Boonvisudhi K., Wang H.G., 1995, Assay For Movement Of Lepidopteran Transposon IFP2 in Insect Cells Using a Baculovirus Genome as a Target DNA, *Virology*, 211, 397-407

Gendrel A., Lippman Z., Yordan C., Colot V., Martienssen R.A., 2002, Dependence of Heterochromatic Histone H3 Methylation Patterns on the Arabidopsis Gene DDM1, *Science*, 297, 1871-1873

Giardine B., Riemer C., Hardison R.C., Burhans R., Elnitski L., Shah P., Zhang Y., Blankenberg D., Albert I., Taylor J., Miller W., Kent W.J., Nekrutenko A., 2005, Galaxy: A Platform for Interactive Large-scale Genome Analysis, *Genome Research*, 15, 1451-1455

Goecks J., Eberhard C., Too T., Nekrutenko A., Taylor J., 2013, Web-based Visual Analysis for High-throughput Genomics, *BMC Genomics*, 14, number 397

Goecks J., Nekrutenko A., Taylor J., The Galaxy Team, 2010, Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible and Transparent Computational Research in the Life Sciences, *Genome Biol.*, 11, 86

Gompel N., 2005, *Drosophila* Germline Transformation, downloaded from http://www.ibdml.univ-mrs.fr/equipes/BP_NG/Methods-files/injection.pdf

Gonzalez J. and Petrov D.A., 2009, The Adaptive Role of Transposable Elements in the *Drosophila* Genome, *Gene*, 448, 124-133

Guil L., Barron M.G., Gonzalez J., 2014, The Transposable Element Bari-Jheh Mediates Oxidative Stress Response in *Drosophila*, *Molec. Eco.*, 23, 2020-2030

Han J.S., 2010, Non-long Terminal Repeat (Non-LTR) retrotransposons: Mechanisms, Recent Developments, and Unanswered Questions, *Mobile DNA*, 1, 15

Hancks D.C. and Kazazian Jr. H.H., 2012, Active Human Retrotransposons: Variation and Disease, *Curr. Op. in Gen. and Dev.*, 22, 191-203

Handler A.M. and Harrell R.A., 1999, Genetic Transformation of *Drosophila melanogaster* with the piggyBac Transposon Vector, *Insect Molec. Biol.*, 8, 449-457

Handler A.M., 2002, Use of the piggyBac Transposon for Germ-line Transformation of Insects, *Insect Biochem. and Molec. Biol.*, 32, 1211-1220

Hedges L., Brownlie J., O'Neill S., Johnson K. 2008, *Wolbachia* and Virus Protection in Insects, *Science*, 322, 702

Holt R.A. *et al*, 2002, The Genome Sequence of the Malaria Mosquito *Anopheles gambiae*, *Science*, 298, 129-149

Honeybee Genome Sequencing Consortium, 2006, Insights into Social Insects from the Genome of the honeybee *Apis mellifera*, *Nat.*, 443, 931-949

Hurst G., Jiggins F. M., Graf von Der Schulenburg J. H., Bertrand D., 1999, Male Killing Wolbachia in Two Species of Insects, *PNAS*, 266, 735-740

International Human Genome Sequencing Consortium, 2001, Initial Sequencing and Analysis of the Human Genome, *Nat.*, 412, 565-566

International Silkworm Genome Consortium, 2008, The Genome of a Lepidopteran Model Insect, the Silkworm *Bombyx mori*, *Insect Biochem. and Molec. Biol.*, 38, 1036-1045

Ioannidis P., Hotopp J.C.D., Sapountzis P., Siozios S., Tsuamis G., Bordenstein S.R., Baldo L., Werren j.H., Bourtzis K., 2007, New Criteria for Selecting the Origin of DNA Replication in Wolbachia and Closely Related Bacteria, *BMC Genom.*, 8, article number 182

Janicki M., Rooke R., Yang G., 2011, Bioinformatics and Genomic Analysis of Transposable Elements in Eukaryotic Genomes, *Chromosome Res.*, 19, 787-808

Johnson L.J. and Brookfield F.Y., 2006, A Test of the Master Gene Hypothesis for Interspersed Repetitive DNA Sequences, *Molec. Biol. Evol.*, 23, 235-239

Jurka J., Kapitonov V.V., Pavlicek A., Klonowski P., Kohany O., Walichiewicz J., 2005, Repbase Update, a Database of Eukaryotic Repetitive Elements, *Cytogenetic and Genome Research*, 110, 462-467

Kaminker J.S., Bergmann C.M., Kronmiller B., Carlson J., Svirskas R., Patel S., Frise E., Wheeler D.A., Lewis S.E., Rubin G.M., Ashburner M., Celniker S.E., 2002, The Transposable Elements of the *Drosophila melanogaster* Euchromatin: a Genomics Perspective, *Genome Biol.*, 3, research0084.1–0084.20

Kass D.H., Batzer M., Deininger P.L., 1995, Gene Conversion as a Secondary Mechanism of Short Interspersed Element (SINE) Evolution, *Molec. Cell Biol.*, 15, 19-25

Kelley L.A. and Sternberg M.J.E., 2009, Protein Structure Prediction on the Web: A Case Study Using the Phyre Server, *Nat. Protocols*, 4, 363-371

Khayrandish A. And Wood R.J., 1993, A Multiple Basis for Insecticide Resistance in a Strain of *Culex quinquefasciatus* (Diptera: Culicidae) from Muheza, Tanzania, Studied as Resistance Declined, *Bulletin of Entomological Research*, 83, 75-86

Kidwell M.G. and Lisch D.R., 2001, Perspective: Transposable Elements, Parasitic DNA and Genome Evolution, *Evo.*, 55, 1-24

Kidwell M.G. and Lisch D.R., 2002, Transposable Elements as Sources of Genomic Variation, in *Mobile DNA II*, edited by Craig N.L., American Society for Microbiology, Washington, D.C. 59-92

Kiszewksi A., Mellinger A., Spielman A., Malaney P., Sachs S.E., Sachs J., 2004, A Global Index Representing the Stability of Malaria Transmission, *American J. of Tropical Medicine and Hygiene*, 70, 486-498

Kyle J.L. and Harris E., 2008, Global Spread and Persistence of Dengue, *Ann. Rev. of Microbiology*, 62, 71-92

Labrador M., Corces V.G., 2002, Interactions between Transposable Elements and the Host Genome, in *Mobile DNA II*, eds Craig N., Craigie R., Gellert M., Lambowitz A., American Society for Microbiology, Washington, D.C., 1008-1039

Laity J.H., Lee B.M., Wright P.E., 2001, Zinc Finger Proteins: New Insights into Structural and Functional Diversity, *Curr. Op. in Structural Biol.*, 11, 39-46

Langer-Safer P.R., Levine M., Ward D.C., 1982, Immunological Method for Mapping Genes on Drosophila Polytene Chromosomes, PNAS, 79, 4381-4385

Langmead B., Salzberg S., 2012, Fast gapped-read alignment with Bowtie 2, *Nat. Methods*, 9, 357-359

Lee Y.C.G. and Langley C.H., 2010, Transposable Elements in Natural Populations of *Drosophila melanogaster*, *Phil. Transactions of the Royal Society B*, 365, 1219-1228

Leonardo T.E. and Nuzhdin S.V., 2002, Intracellular Battlegrounds: Conflict and Cooperation between Transposable Elements, *Genetical Research*, 80, 155-161

Li X., Lobo N., Bauser C.A., Fraser Jr. M.J., 2001, The Minimum Internal and External Sequence Requirements for Transposition of the Eukaryotic Transformation Vector *piggyBac*, *Molec. Genet. Genom.*, 266, 190-198

Lidholm D.A., Lohe A.R., Hartl D.L., 1993, The Transposable Element Mariner Mediates Germline Transformation in *Drosophila melanogaster*, *Gen.*, 134, 859-868

Lippman Z., Gendrel A., Black M., Vaughn M.W., Dedhia N., McCombie W.R., Lavine K.m Mittal V., May B., Kasschau K.D., Carrington J.C., Doerge R.W., Colot V., Martienssen R., 2004, Role of Transposable Elements in Heterochromatin and Epigenetic Control, *Nat.*, 430, 471-476

Lobo N.F., Hua-Van A., Nolen B.M., Fraser Jr M.J., 2002, Germline Transformation of the Yellow Fever Mosquito, *Aedes aegypti*, Mediated by Transpositional Insertion of a piggyBac vector, *Insect Molec. Biol.*, 11, 133-139

Loukeris T.G., Livadaras I., Arca B., Zabalou S., Savakis C., 1995, Gene Transferinto the Medfly, *Ceratitis capitata*, with a *Drosophila hydei* Transposable Element, *Science*, 170, 2002-2005

Manoharan M., Chong M.N.F., Vaïtinadapoulé A., Frumence E., Sowdhamini R., Offmann B., 2013, Comparative Genomics of Odorant Binding Proteins in *Anopheles gambiae*, *Aedes aegypti*, and *Culex quinquefasciatus, Genome Biol. and Evo., 5, 163-180*

Marinoti O., Cerqueira G.C., de Almeida L.G.P., Ferro M.I.T., de Silva Loreto E.L. *et al*, 2013, The Genome of *Anopheles darlingi*, the Main Neotropical Malaria Vector, *Nucleic Acids Research*, 41, 7387-7400

Marshall J.M. and Taylor C.E., 2009, Malaria Control with Transgenic Mosquitoes, *PLOS Med.*, 6, 164-168

Martens J.H.A., O'Sullivan R.J., Braunschweig U., Opravil S., Radolf M., Steinlein P., Jenuwein T., 2005, The Profile of Repeat-Associated Histone Lysine Methylation States in the Mouse Epigenome, *EMBO*, 24, 800-812

Marunouchi T. and Hosoya H., 1984, Isolation of an Autonomously Replicating Sequence (ARS) from Satellite DNA of *Drosophila melanogaster*, *Molec. Genet. Genom.*, 196, 258-265

Maruyama K. and Hartl D.L., 1991, Evolution of the Transposable Element Mariner in Drosophila Species, *Gen.*, 128, 319-329

Mason J.M., Frydrychova R.C., Biessmann H., 2007, *Drosophila* Telomeres: An Exception Providing New Insights, *BioEssays*, 30, 25-37

McClintock B., 1951, Chromosome Organization and genic expression, Cold Spring Harbor Symposia, *Quant. Biol.*, 16, 13–47

McClintock B., 1956, Controlling Elements and the Gene, Cold Spring Harbor Symposia, *Quant. Biol.*, 21, 197-216

McGraw E.A. and O'Neill S.L., 2013, Beyond Insecticides: New Thinking of an Ancient Problem, *Nat. Rev. Micobio.*, 11, 181-193

Megy, K. *et al*, 2012, VectorBase: Improvements to a Bioinformatics Resource for Invertebrate Vector Genomics, *Nucleic Acids Research*, 40: D729-734

Metaxakis A., Oehler S., Klinakis A., Savakis C., 2005, Minos as a Genetic and Genomic Tool in *Drosophila melanogaster*, *Gen.*, 171, 571-581

Metcalfe C.J. and Casane D., 2014, Modular Organisation and Reticulate Evolution of the ORF1 of Jockey Superfamily Transposable Elements, *Mobile DNA*, 5, article number 19

Moran J.V. and Gilbert N., 2002, Mammalian LINE-1 Retrotransposons and Related Elements, in *Mobile DNA II*, eds. Craig N., Craigie R., Gellert M., Lambowitz A., American Society for Microbiology, 836-869

Morgan H.D., Sutherland H.G., Martin D.I., Whitelaw E., 1999, Epigenetic Inheritance at the Agouti Locus in the Mouse, *Nat. Rev. Gen.,* 23, 314–318

Morgulis A., Coulouris G., Raytselis Y., Madden T.L., Agarwala R., Schäffer A.A., 2008, Database Indexing for Production MegaBLAST Searches, *Bioinfo.*, 24, 1757-1764

Mouches C., Bensaadi N., Salvado J.C., 1992, Characterisation of a LINE Retroposon Dispersed in the Genome of Three Non-sibling *Aedes* Mosquito Species, *Gene*, 2, 183-190

Nene V. *et al*, 2007, Genome Sequence of *Aedes Aegypti*, a Major Arbovirus Vector, *Science*, 316, 1718-1723

Noe L. and Kucherov G., 2005, YASS: Enhancing the Sensitivity of DNA Similarity Search, *Nucleic Acids Research*, 33, 540-543

Nouaud D., Boeda B., Levy L., Anxolabehere D., 1999, A P Element has Induced Intron Formation in *Drosophila*, *Molec. Biol. Evo.*, 16, 1503-1510

O'Brochta D.A. and Handler A.M., 1988, Mobility of P Elements in Drosophilids and Nondrosophilids, *PNAS*, 85, 6052-6056

O'Brochta D.A., Sethuraman N., Wilson R., Hice R.H., Pinkerton A.C., Levesque C.S., Bideshi D.K., Jasinskiene N., Coates C.J., James A.A., Lehane M.J., Atkinson P.W., 2003, Gene Vector and Transposable Element Behaviour in Mosquitoes, *J. Exp. Biol.* 206, 3823-3834

Oliver K.R. and Greene W.K., 2009, Transposable Elements: Powerful Facilitators of Evolution, *Bioessays*, 31, 703-714

Osanai F.M., Suetsugu Y.M., Mita K., Fujiwara H., 2008, Genome-wide Screening and Characterisation of Transposable Elements and Their Distribution Analysis in the Silkworm, *Bombyx mori*, *Insect Biochem. and Molec. Biol.*, 38, 1046-1057

Patrushev L.I. and Minkevich I.G., 2008, The Problem of the Eukaryotic Genome Size, *Biochem.*, 73, 1519-1552

Petrov D.A., Schutzman J.L., Hartl D.L., Lozoyskaya E.R., 1995, Diverse Transposable Elements are Mobilized in Hybrid Dysgenesis in *Drosophila virilis*, *PNAS*, 92, 8050-8054

Rai K.S., 2010, Insights from Mosquito Evolution: Patterns, Tempo and Speciation in *Nature at Work: Ongoing Saga of Evolution*, ed Sharma V.P., National Academy of Sciences, India, 197-218

Ray D.A., Feschotte C., Pagan H.J.T., Smith J.D., Pritham E.J., Arensburger P., Atkinson P.W., Craig N.L., 2008, Multiple Waves of Recent DNA Transposon Activity in the Bat, *Myotis lucifugus*, *Genome Research*, 18, 717-728

Raymond M., Beyssat-Arnaouty V., Sivasubramanian N., Mouches C., Georghiou G.P., Pasteur N., 1989, Amplification of Various Esterase B's Responsible for Organophoshate Resistance in *Culex* mosquitoes, *Biochem. Gen.*, 27, 417-423

Razin S.V., Borunova V.V., Maksimenko O.G., Kantidze O.L., 2012, Cys_2Hys_2 Zinc Finger Protein Family: Classification, Functions and Major Members, *Biochem.*, 77, 217-226

RepRap, 2013, RepRap, last edited 13 October 2013, http://reprap.org/wiki/Main_Page

Rio D.C., 1990, Molecular Mechanisms Regulating *Drosophila* P element Transposition, *Annual Reviews Genetics*, 24, 543–578
Rio D.C., 2002, P Transposable Elements in *Drosophila melanogaster*, in *Mobile DNA II*, eds Craig N., Craigie R., Gellert M., Lambowitz A., Washington, D.C., 484-518

Rozen S. and Skaletsky H.J., 2000, 'Primer3 on the WWW for General Users and for Biologist Programmers', in Krawetz S., Misener S. (eds), *Bioinfo. Methods and Protocols: Methods in Molec. Biol.*, Humana Press, Totowa, NJ, 365-386

Rubin G.M. and Spradling A.C., 1982, Genetic Transformation of *Drosophila* with Transposable Element Vectors, *Science*, 218, 348-353

Saito K., Nishida K.M., Mori T., Kawamura Y., Miyoshi K., Nagami T., Siomi H., Siomi M.C., 2006, Specific Association of Piwi with rasiRNAs Derived from Retrotransposon and Heterochromatic Regions in the *Drosophila* Genome, *Gen. and Dev.*, 20, 2214-2222

Santamaria P., 1986, Injecting Eggs, in *Drosophila: A Practical Approach*, eds. Roberts D.B., Information Printing Ltd., Oxford, England

Schwartz A., Chan D.C., Brown L.G., Alagappan R., Pettay D., Disteche C., McGillivray B., de la Chapelle A., Page D.C., 1998, Reconstructing Hominid Y Evolution: X-homologous Block, Created by X-Y Transposition, was Disrupted by Yp Inversion through LINE-LINE Recombination, *Human Molec. Genet.*, 7, 10-11

Shpiz S., Kwon D., Uneva A., Kim M., Klenov M., Rozovsky Y., Georgiev P., Savitsky M., Kalmykova A., 2007, Characterisation of *Drosophila* Telomeric Retroelement TAHRE: Transcription, Transpositions and RNAi-based Regulation of Expression, *Molec. Biol. and Evo.*, 24, 2535-2545

Sijen T. and Plasterk R.H.A., 2003, Transposon Silencing in the *Caenorhabditis elegans* Germ Line by Natural RNAi, *Nature, 426, 310-314*

Simmons M.J., Ryzek D., Lamour C., Goodman J.W., Kummer N.E., Merriman P.J., Cytotype Regulation by Telomeric P Elements in *Drosophila* melanogaster: Evidence of Involvement of an RNA Interference Gene, *Genet.*, 176, 1945-1955

Sinkins S.P. and Gould F., 2006, Gene Drive Systems for Insect Disease Vectors, *Nat. Rev. Gen.*, 7, 427-435

Slotkin R.K. and Martienssen R., 2007, Transposable Elements and the Epigenetic Regulation of the Genome, *Nat. Rev. Gen.*, 8, 272-285

Smit A.F.A., Hubley R., Green P., unpublished data, Current Version: Open-4.05 (RMLib: 20140131 & Dfam: 1.2)

Smith C.D., Zimin A., Holt C., Abouheif E., Benton R., *et al*, 2011, Draft Genome of the Globally Widespread and Invasive Argentine Ant (*Linepithema humile*), *PNAS*, 108, 5673-5678

Smith C.R., Smith C.D., Robertson H.M., Helmkampf M., Zimin A., *et al*, 2011, Draft Genome of The Red Harvester Ant *Pogonomyrmex barbatus*, *PNAS*, 108, 5667-5672

Spradling A. and Rubin G.M., 1982, Transposition of Cloned P Elements into *Drosophila* Germ Line Chromosomes, *Science*, 218, 341-347

Spradling A.C., Stern D., Beaton A., Rhem E.J., Laverty T., Mozden N., Misra S., Rubin G.M., 1999, The Berkeley *Drosophila* Genome Project Gene Disruption Project: Single P Element Insertions Mutating 25% of Vital *Drosophila* Genes, *Gen*., 153, 135-177

Suen G., Teiling C, Li L, Holt C, Abouheif E, *et al*, 2011, The Genome Sequence of the Leaf Cutter Ant *Atta cephalotes* Reveals Insights Into Its Symbiotic Lifestyle, *PLoS Gen*., 7, e1002007

Talamas E., Jackson L., Koeberl M., Jackson T., McElwee J.L., Hawes N.L., Chang B., Jablonski M.M., Sidjanin D.J., 2006, Early Transposable Element Insertion in Intron 9 of the *Hsf4* Gene Results in Autosomal Recessive Cataracts in *lop11* and *ldis1* Mice, *Genomics*, 88, 44-51

Tautz D., and Domazet-Loso T., 2011, The Evolutionary Origin of Orphan Genes, *Nat. Rev. Gen*., 12, 692-702

The International Aphid Genomics Consortium, 2010, Genome Sequence of the Pea Aphid *Acyrthosiphon pisum*, *PLOS Biol*., 8, Article number e1000313

The R Core Team, 2015, R Version 3.1.3 in R: A Language and Environment for Statistical Computing from http://www.r-project.org/

Touriol C., Bornes S., Bonnal S., Audigier S., Prats H., Prats A., Vagner S., 2003, Generation of Protein Isoform Diversity by Alternative Initiation of Translation at Non-AUG codons, *Bio. Cell*, 95, 169-178

Tribolium Genome Sequencing Consortium, 2008, The Genome of the Model Beetle and Pest *Tribolium castaneum*, *Nat*., 452, 949-955

Tu Z. and Coates C., 2004, Mosquito Transposable Elements, *Insect Biochem. and Molec. Biol*., 34, 631-644

Vagin V.V., Sigova A., Li C., Seitz H., Gvozdev V., Zamore P.D., 2006, A Distinct Small RNA Pathway Silences Selfish Genetic Elements in the Germline, *Science*, 313, 320-324

Warren W.D., Atkinson P.W., O'Brochta D.A., 1994, The Hermes Transposable Element from the House Fly, Musca domestica, is a Short Inverted Repeat-Type Element of the Hobo, Ac and Tam3 (hAT) Element Family, *Gen. Research*, 64, 87-97

Waterhouse R.M., Wyder S., Zdobnov E.M., 2008, The *Aedes Aegypti* genome: a Comparative Perspective, *Insect Molec. Biol.,* 17, 379-396

Weber B. and Schmidt T., 2009, Nested Ty3-gypsy Retrotransposons of a Single *Beta Procumbens* Centromere Contain a Putative Chromodomain, *Chromosome Research*, 17, 379-396

Wicker T., Sabot F., Hua-Van A., Bennetzen J.L., Capy P., Chalhoub B., Flavell A., Leroy P., Morgante M., Panaud O., Paux E., Sanmiguel P., Schulman A.H., 2007, A Unified Classification System for Eukaryotic Transposable Elements, *Nat. Rev. Gen.*, 8, 973-982.

Wimmer E.A., 2003, Applications of Insect Transgenesis, *Nat. Rev. Gen*., 4, 225-232

Zhang Z., Schwartz S., Wagner L., Miller W., 2000, A Greedy Algorithm For Aligning DNA Sequences, *J. Comput. Biol*., 7, 203-214