THE EVOLUTION OF TRANSPOSABLE ELEMENTS IN HUMANS AND *DROSOPHILA*

Pamela Styles, BSc.

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

April 2010

Abstract

The different genomic environments in which transposable elements reside in the Great Apes and the *Drosophila* result in substantial differences between the evolution of transposable elements in these two groups of organisms. In the Great Apes, where deletion of transposable elements is relatively rare, elements tend to be retained in the genome to the extent that complete sets of elements belonging to a particular transposable element family can be obtained. In *Drosophila*, there is a rapid turnover of transposable elements, imposing strong selection pressure on transposable elements to be able to infect new hosts. This study investigates the evolution of transposable elements in these two genomic environments.

Complete sets of elements belonging to young Alu subfamilies in humans and closely-related species are used to investigate factors involved in their evolutionary history, such as mutation and gene conversion. The application of the master gene model, and other proposed models of the proliferation of young Alu subfamilies, are considered in light of the results obtained. The evolution of the AluYg, Yh and Yi lineages are investigated using a C++ program to simulate their evolutionary history. The results of the simulations are compared to statistics such as theta and pi, as well as the number of shared mutations and the proliferation time, in order to determine possible, and likely, values for parameters such as the retrotransposition rate and the number of source elements for each subfamily. The results suggest that although the master gene model may apply to some lineages, it is not the best model to explain the evolutionary history of all young Alu subfamilies.

The selection pressure on transposable elements in *Drosophila* results in a high level of horizontal transfer of these elements among species of the *Drosophila* genus. In this study, the twelve sequenced *Drosophila* genomes are used to investigate the frequency of horizontal transfer within these twelve species using a large dataset of transposable element sequences from the DNA transposons, as well as LTR, and non-LTR, retrotransposons. Horizontal transfer is inferred where identity between transposable elements of the same family in different species exceeds that between the coding regions of the *Adh* gene in the relevant species. Cases are further supported by evidence from the distribution of the transposable element family across the *Drosophila* genus, and phylogenetic incongruence, which in many cases elucidates likely directions of transfer. The results suggest that horizontal transfer may be even more common than previously thought, and appears to be most common for the LTR retrotransposons. The possibility that possession of the *env* gene may result in higher rates of horizontal transfer of LTR retrotransposons is investigated, and the *env* open reading frame is found to be under selective constraint.

Acknowledgements

I would first and foremost like to thank my wonderful husband David Herrington. I don't see how I could have done this without you. Thank you for your never-ending support, encouragement, love and Excel expertise! I'm sure you've learnt more about transposable elements than you ever wanted to know. I'm really grateful for all your time and patience, you've helped me so much, and I hope I can do the same for you over the next few years.

Thank you to John Brookfield, my brilliant supervisor, for all your input into the project and your years of help and guidance. I know that I've been very lucky to have you as a supervisor, and I'm not sure I would have been able to do this with anyone else.

I'd like to acknowledge the input of Bob Scarle, who designed the program interface in chapter 3, John Armour for suggesting the method used to determine the polymorphism status of Alu subfamilies in chapter 2, and Sophie Austin for contributions to the work on Nobel in chapter 5. I'd also like to thank everyone else who has given me feedback over the years, especially my second supervisor Chris Wade.

Thanks to old members of the JFYB lab - Davide van Vlijmen (for your programming assistance and keeping the dark old office entertaining!) and Dr. Rachel Clark, my good friend, it wouldn't have been the same without you.

Everyone in B134 and B135 - now I know what it's like to work in an office with a window! Thank you for all the entertaining discussions, tea breaks, and making finishing my research a lot easier than it would have been stuck in the dark little office on my own!

Amy Hall, thank you for your friendship and more cups of tea than I count! You've been there when I needed someone to talk to, and have ultimately made getting the final stages of my PhD done possible!

Mom and Dad, thank you for your continued support and always being there to listen and guide me, and for trying to understand what I've been talking about for the past three years! You're the best parents anyone could ask for and I know I wouldn't be where I am today without you.

And finally I'd like to thank Java, Perl and Bell for keeping me company over the past few years. You were fantastic little friends and I miss having you around.

Table of Contents

SUPPLEMENTARY DATA CD

List of Abbreviations

A	Adenine
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like Alignment Tool
bp	Base Pairs
C	Cytosine
cDNA	Complementary DNA
DDT	Dichlorodiphenyltrichloroethane
DNA	Deoxyribonucleic acid
FAM	Free Alu Monomer
FLAM	Free Left Alu Monomer
FRAM	Free Right Alu Monomer
G	Guanine
HIV	Human Immunodeficiency Virus
IPL	Insertion Polymorphism Level
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
mRNA	Messenger RNA
mya	Million Years Ago
myr	Million Years
NCBI	National Center for Biotechnology Information
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
RNA	Ribonucleic acid
RT	Reverse Transcriptase
SINE	Short Interspersed Nuclear Element
SNP	Single Nucleotide Polymorphism
SRP	Signal Recognition Particle
T	Thymine
TIR	Terminal Inverted Repeat
TPRT	Target-Primed Reverse Transcription
tRNA	Transfer RNA
U	Uracil
UTR	Untranslated Region
VLP	Virus-like Particle

Table of Figures

Table of Tables

Chapter 1 - Introduction

1.1 Types of transposable element

Transposable elements are DNA sequences that are able to move to new genomic locations within a cell, by the process of transposition. They were first discovered in the 1950s by Barbara McClintock during her work on maize (McClintock 1953). Transposable elements comprise a high percentage of the genomes of higher eukaryotes, but are also found in organisms which are generally thought to have very compact genomes. These include bacteria such as the Mycobacteria (Bull et al. 2003) and the Enterobacteria (Bachellier et al. 1999), and the yeast *Saccharomyces cerevisiae*. Around 50% of mammalian genomes are derived from transposable elements, but this is significantly higher in plants, where around 80% of some genomes are transposon-derived (van de Lagemaat et al. 2005). As a consequence of their high copy number and mobility, transposable elements have played an important role in the evolution of many eukaryotic genomes, having both deleterious and advantageous effects. For example, ectopic recombination between similar or identical repeats can result in deletion of essential host sequences, or transposable element integration can interrupt a host gene or regulatory region (Charlesworth et al. 1994). Conversely, many gene regulatory regions in eukaryotes are derived from transposable elements (van de Lagemaat et al. 2005), with nearly 25% of human promoters containing sequences derived from transposable elements (Jordan et al. 2003). Helitrons, a type of transposable element present in eukaryotes, are capable of capturing host genes, which may result in domain shuffling and the evolution of novel proteins (Kapitonov and Jurka 2007). Transposable elements have also become domesticated, or exapted, to perform advantageous functions for their host (Gombart et al. 2009;Gonzalez et al. 2009;Kidwell and Lisch 2001).

There are two main classes of transposable element, class I and class II, which differ in their mechanism of transposition. Class II elements, or DNA transposons, mobilise by a non-replicative “cut-and-paste” mechanism. During transposition, class II elements are excised from the chromosome then reintegrate at a new locus, such that the integration event is not automatically associated with an increase in copy number of the transposable element. Transposable elements can be either autonomous or non-autonomous. An element which is capable of mediating its own transposition is defined as autonomous. A non-autonomous element is one which either does not possess sequences encoding the enzymes required for transposition, or in which these sequences are present but are mutated. Non-autonomous elements possess intact enzyme recognition sites, such that the element can be mobilised by enzymes produced by other sequences. Further to these two types of elements, some will possess mutations within the enzyme recognition sites, and therefore not be capable of transposition. This strict definition creates a clear dichotomy between these two types of elements, as a single point mutation within an open reading frame of an autonomous element can render the element non-autonomous. The activity of various autonomous elements may, however, vary. There may also be some involvement of host proteins, such as DNA ligase, in the mobilisation of autonomous elements.

The enzyme responsible for transposition of DNA transposons is called transposase, and is generally encoded by the DNA transposon itself, although non-autonomous elements make use of transposase encoded by other transposons. Transposase binds near the inverted repeats which flank a transposon (Figure 1.1), and to target DNA. The two strands of the target are then cut, at staggered sites, which leads to the generation of target site duplications (Kazazian 2004). The structure of a DNA transposon typically includes terminal inverted repeats at the 5' and 3' ends, which are essential for transposition. The internal portion of autonomous elements contains the open reading frame for the transposase enzyme. There are several types of DNA transposon, including Politrons, Helitrons, and self-synthesising elements (Kapitonov and Jurka 2007). Helitrons, which will be discussed in chapter 4, have been identified across many eukaryotic species, and replicate differently from the majority of DNA transposons, following a rolling circle mechanism of replication (Kapitonov and Jurka 2007).

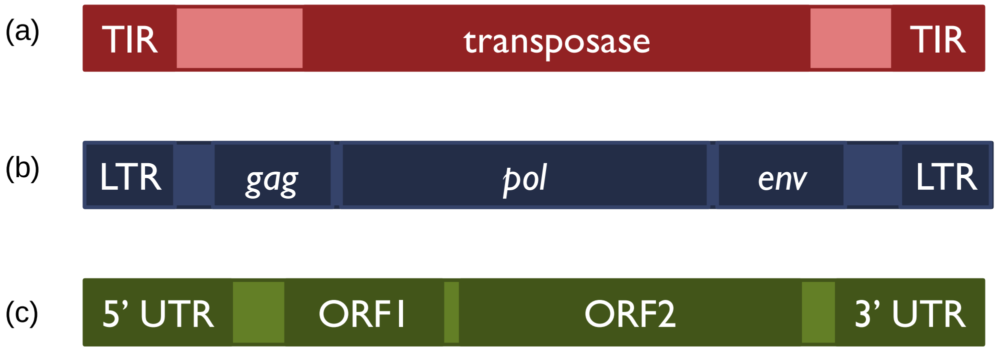


Figure 1.1: The three main types of transposable element, (a) DNA transposon, (b) LTR retrotransposon and (c) non-LTR retrotransposon.

Class I elements, or retrotransposons, mobilise by a replicative “copy-and-paste” mechanism, whereby integration at a new locus results in an increase in element copy number. The element is first transcribed, from an internal TATA-less promoter (Arkhipova 1995), into RNA, which is then reverse transcribed into a cDNA copy, which then integrates at a new genomic location. Retrotransposons can be classified as LTR- or non-LTR retrotransposons, which refers to the presence, or absence, of long terminal repeats at their 5’ and 3’ ends. Non-LTR retrotransposons possess a poly(A) tail at the 3' end, which forms during integration, and therefore varies in length between transposable elements of the same family.

LTR-retrotransposons resemble retroviruses in terms of their genomic structure. This family of transposable elements includes the endogenous retroviruses, which have a lifecycle identical to that of a retrovirus, except that the new genomes produced are not generally assembled into a viral particle which can then infect other cells. Instead, these elements, like all transposable elements, propagate within a single cell. Consequently, only transposition events which occur within cells of the germline have the potential to have an impact on evolution. Endogenous retroviral genomes have either a defective or totally absent *env* gene, which would encode components of the viral particle. It is for this reason that they cannot exit the cell. However, these elements do carry functional *gag* and *pol* genes. The *gag* gene encodes structural proteins, whereas the *pol* gene encodes the reverse transcriptase, cleavage protease and integrase functions required for retrotransposition. Reverse transcription occurs within a virus-like particle in the cytoplasm (Kazazian 2004). Some LTR retrotransposons do possess a functional *env* gene, which encodes the virus-like particle, and will be discussed in detail in chapter 5.

Non-LTR retrotransposons do not possess long terminal repeats. They have been shown to fall into at least eleven clades, dating back to the pre-Cambrian (Malik et al. 1999). Long Interspersed Nuclear Elements (LINEs), which are approximately 3-5kb in length, are an example of an autonomous family of non-LTR retrotransposons. This family includes the L1 element in mammals, which is the only active mammalian autonomous LINE. The vast majority of L1 elements are truncated at their 5’ ends, however, those that are intact carry two open reading frames (ORFs), which encode proteins designated as ORF1p and ORF2p, separated by 63nt of non-coding DNA. ORF1p is a nucleic acid binding protein, which has been shown to have a possible chaperone function (Dewannieux and Heidmann 2005). ORF2p encodes the endonuclease and reverse transcriptase (RT) activities which are essential for retrotransposition. The endonuclease domain cleaves a specific 5bp target site in the DNA at the integration site, allowing the cDNA copy of the element produced by the RT domain, via the process of target-primed reverse transcription (TPRT) to insert (Batzer and Deininger 2002). L1 elements comprise approximately 17% of the human genome by mass, reaching over 500,000 copies in the past 150myr (Lander et al. 2001).

Transposable elements can be classified into families and superfamilies. The investigation of the transposable elements of the human genome in this study focuses on Alu elements, a type of non-LTR retrotransposon. In *Drosophila*, a broad range of transposable elements from both class I and class II are investigated. These are classified into superfamilies, such as Tc1/mariner and Gypsy. Within each superfamily are many individual transposable element families.

1.2 Alu Elements

Short Interspersed Nuclear Elements (SINEs) are an example of non-autonomous non-LTR retrotransposons. These are generally around 100-800bp of non-coding DNA, which include, for example, the ~300bp Alu element, which utilises the L1 ORF2p enzyme for its retrotransposition. Alu elements do not require L1 ORF1p in order to mobilise (Dewannieux et al. 2003). Alu elements are so named due to the presence of an *Alu*I restriction enzyme site in their sequence. They are derived from the 7SL RNA gene, which produces a component of the Signal Recognition Particle (SRP) ribonucleoprotein. SRP interacts with ribosomes, and mediates the movement of nascent proteins across cell membranes (Maity et al. 2006).

Alu elements are not the only SINE derived from SRP RNA, and retrotransposons of this nature are found in several orders across the Euarchontoglires (Primates, Rodents and Scandentians). In rodents, these elements belong to the B1 family, and in Scandentians, to the Tu family (Vassetzky et al. 2003). In primates, in addition to Alu elements, ancestral forms are seen, for example the Free Alu Monomer (FAM), and the Free Left and Right Alu Monomers (FLAM and FRAM, respectively). A SINE has also been identified containing *AluI* restriction sites in *Amphioxus*, however, this SINE is derived from a tRNA and is not related to mammalian 7SL-derived SINEs (Holland 2006). Previously, it was believed that FAM was the common ancestor of the 7SL-derived SINEs in the Supraprimates, however, evidence has been recently presented that suggests that FAM is restricted to primates, and that FLAM subtype A (FLAM-A), known as proto-B1 (PB1) in rodents, is the true common ancestor (Kriegs et al. 2007).

Alu elements are a dimeric fusion of FLAM-C and FRAM, and contain an internal RNA polymerase III promoter in the left half which is essential for their transcription and subsequent retrotransposition. The internal promoter in the right half has been inactivated by mutations (Li and Schmid 2004). Promoter box A is found between positions 6 and 15, and box B between positions 75 and 84 (Figure 1.2). The left half also contains a binding site for the SRP 9/14 heterodimer (box 1 – positions 15-17, box 2 – positions 22-31) (Aleman et al. 2000). As this protein associates with ribosomes, it has been suggested that the interaction of Alu RNA with SRP9/14 might be what brings it into contact with L1 proteins as they are being translated, allowing it to compete effectively for their activity (Dewannieux et al. 2003;Li and Schmid 2004). This is supported by the tight coevolution between SRP9/14 in primates and the Alu consensus sequence (Li and Schmid 2004). However, Alu RNA binding to SRP 9/14 has decreased throughout primate evolution (Kazazian 2004).

GGCCG**GGCGCGGTGGCT**CACG**CCTGTAATCC**CAGCACTTTGGGAGGCCGAGGCGGGAGGATTG
CTTGAGCCAGGA**GTTCGAGACC**AGCCTGGGCAACATAGCGAGACCCCGTCTCTACAAAAATA
CAAAAATTAGCCGGGCGTGGTGGCGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAG
GAGGATCGCTTGAGCCCAGGAGTTCGAGGCTGCAGTGAGCTATGATCGCGCCACTGCACTCCA
GCCTGGGCGACAGAGCGAGACCCTGTCTC

Figure 1.2: The general Alu consensus sequence. The A and B boxes of the internal RNA polymerase III promoter are shown in red. The SRP9/14 binding site boxes 1 and 2 are underlined in blue.

Alu elements have variable-length poly(A) tails at their 3’ ends. These are unstable and are therefore variable not only between Alu elements at different loci, but also between individuals when looking at a single locus. These poly(A) tails may be involved in stability of Alu RNA and are likely involved in recognition of the RNA by L1 machinery (Dewannieux and Heidmann 2005), as L1 elements also possess poly(A) tails. Efficient retrotranspositions in a model system are only seen if the Alu element has a poly(A) tail, rather than a poly(C), poly(G) or poly(T) tract (Dewannieux and Heidmann 2005). It is also suggested that a poly(A) tail at least 40 residues in length is required for effective retrotransposition (Johanning et al. 2003).

Alu elements are primate-specific retrotransposons, with some of the youngest subfamilies only present in the human genome. They began amplifying approximately 65 million years ago, following the dimerisation event. They underwent a rapid amplification around 40 million years ago, which corresponds with a burst in processed pseudogene formation (Johanning et al. 2003). Alu elements fall into subfamilies of different ages, which are distinguished by specific diagnostic mutations. These mutations correspond to mutations present in the source gene that gave rise to the subfamily (see below). Alu element subfamilies are named according to standard nomenclature, and fall into three major groups of decreasing age: AluJ, AluS and AluY. Each of these groups, or families, possesses unique diagnostic mutations, which were present in the source elements that were primarily functional at different times in the past (see below). Subfamilies of these groups are designated by letters and numbers, for example the AluYa5 subfamily falls on the “a” lineage derived from AluY, and has 5 diagnostic changes relative to the AluY consensus sequence. As the sequence of an Alu source gene changes over time, daughter elements possessing different shared mutations will be produced. Where copy number is high enough for a group of elements sharing particular mutations to be observed, this group is designated an Alu subfamily. Although this is part of the continuous variation observed among Alu elements, it is the radically differing levels of activity among elements that makes it possible for such subfamilies to be defined.

There are 36 Alu subfamilies listed on Repbase (Jurka et al. 2005), and several of these have been analysed in considerable detail. These include the Alu Ye lineage, consisting of Ye4, Ye5 and Ye6 elements, which have been found to be between 9 and 14 million years old (Salem et al. 2005). Some elements along this lineage are unique to humans, whereas older elements are found in all apes. Of the three subfamilies belonging to this lineage, only Ye5 has established a relatively high copy number.

The AluJ subfamily was most actively transposing around 60mya, AluS 20-60mya (Kapitonov and Jurka 1996), and AluY is still actively transposing. Although the vast majority of older Alus are inactive, some recent retrotransposition by the AluSx subfamily has been observed (Johanning et al. 2003). AluSx, which was actively retrotransposing at high levels during the peak of Alu activity, accounts for 50% of the total number of Alu elements (An et al. 2004).

Alu elements are the most successful SINE in humans, comprising approximately 11% of the genome by mass in over one million copies (Lander et al. 2001). They are densely distributed, with an element occurring on average once every 3000bp (Zhi 2007). Many of the young AluY subfamilies have been shown to be currently retrotransposing, and the current Alu retrotransposition rate has been estimated at approximately 1 in 20 births (Cordaux et al. 2006a). However, this is suggested to be only 1% of the maximum rate of amplification which occurred in the past (An et al. 2004). The rate of transposition is much lower in non-human primates, and this has been shown to be due to an increase in retrotransposition rate along the human lineage rather than a decrease along the others (Mills et al. 2006).

1.3 Gene Conversion

Alu elements, probably as a consequence of their high copy number, undergo relatively frequent gene conversion events (Figure 1.3). Almost 20% of young Alus have undergone partial gene conversion events (Batzer and Deininger

2002). Gene conversion is a non-reciprocal recombination process, whereby resolution of the Holliday junction leads to an unequal crossover. During a gene conversion event, one sequence, the “target” or “acceptor”, is converted by a similar sequence, the “template” or “donor”, such that the acceptor is identical to the donor. Alu elements from different subfamilies are similar enough to each other to act as templates for gene conversion, and examples of both forward (where an older element is converted to a young one) (Salem et al. 2003a) and backward (where a younger element is converted to an old one) have been identified (Roy et al. 2000). Complete gene conversion events, where an Alu is replaced entirely by the sequence of the template, can be identified in instances where, for example, an Alu element from a young subfamily is present in the human genome, and an older Alu is present in the chimpanzee genome at the orthologous locus. Partial gene conversion events have also been identified, where gene conversion tracts of generally around 50-100bp have formed and converted a short section of an Alu (Batzer and Deininger 2002;Roy-Engel et al. 2002). Gene conversion events have contributed considerably to the heterogeneity seen among members of an Alu subfamily.

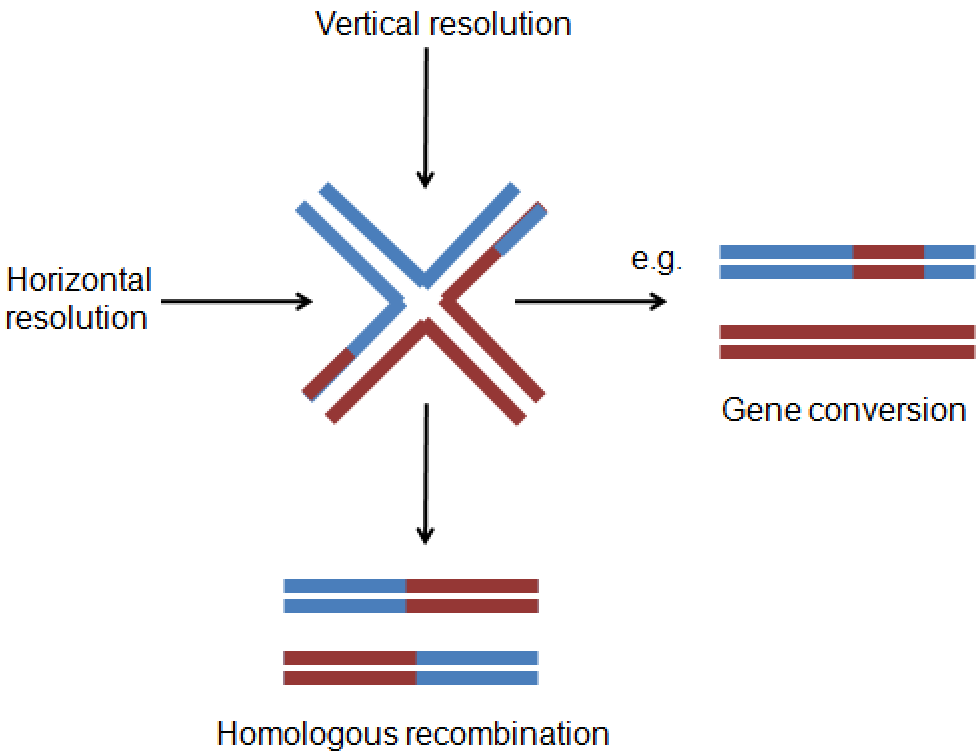


Figure 1.3: Resolution of a Holliday junction leading to either homologous recombination or gene conversion. The Holliday junction intermediate can be resolved either horizontally or vertically. Here, horizontal resolution would result in gene conversion.

The alternative resolution of the Holliday junction formed by two Alu elements is homologous recombination. This process generates duplications and deletions of not only the Alu elements themselves, but occasionally additional flanking genomic DNA. Instances of precise deletion of Alu elements by recombination have been reported (Belle et al. 2005;van de Lagemaat et al. 2005), but these appear to be quite rare (although they may be more frequent in the chimpanzee, (Belle et al. 2005)), and more often than not genomic sequences are deleted, which can lead to human disease. This property of Alu elements gives them the potential to have deleterious effects on the host, and has been suggested to be a factor in driving the diversification of Alus, and consequently the generation of the numerous subfamilies. It is suggested that hosts may have a “carrying capacity” of Alu elements of a certain type, and above this threshold the risk of homologous recombination becomes too high to be maintained, and selection against this occurs at the level of the host. One model of transposable element diversity that has been proposed is based on Lotka-Volterra predator-prey interactions (Abrusan and Krambeck 2006a). The authors suggest that transposable element diversity should be considered in an ecological framework in order to explain the patterns seen.

The general pattern of replacement of older active subfamilies with younger ones generates a phylogeny which generally appears to follow a master gene model (Shen et al. 1991). This will be discussed in more detail below. Such a theory might suggest that homologous recombination and gene conversion are much more likely between highly similar templates, which is a logical assumption. However, gene conversion events have been shown to occur, for example, between divergent AluS sequences and non-divergent AluY sequences. It is possible, however, that this observation is simply seen because it is easier to observe. A gene conversion between an older Alu and a young one can be detected where the old Alu is present in the chimpanzee and the young Alu is present at the orthologous locus in the human genome. Where gene conversion events have, presumably, occurred along the human lineage between two AluY derivatives, an unfilled site would likely be observed in the chimpanzee. This gives no clear way of identifying the sequence as the product of a gene conversion event. The sequence may be suspected as being the product of partial gene conversion if it contains diagnostic mutations, or if the mutations it carries appear to be clustered. However, if highly similar templates are being used, it is possible that very few mutations would be copied by the process, and therefore perhaps only one change, at a diagnostic position, would be seen. This could then subsequently be interpreted as back mutation, rather than necessarily inferring gene conversion.

The similarity of two Alu elements may not be the most important factor in determining the relative likelihood of the pair undergoing a gene conversion event. If gene conversion has occurred, donors and acceptors would be expected to share more mutations than a random pair of Alu elements, if diagnostic mutations are excluded. It has been shown that there is a significant increase in the number of shared mutations between adjacent Alus on chromosome 22 (Zhi 2007), suggesting that proximity is an important factor influencing the probability of gene conversion. This effect was shown to diminish the further apart the two elements are, becoming indistinguishable after 5000bp. However, this study does not look at gene conversion *between* chromosomes. As gene conversion involves an alternative resolution of the Holliday junction to homologous recombination, it seems likely that gene conversion could occur over any range that recombination could occur.

1.4 Insertion Polymorphism

Many insertions of the human-specific AluY subfamilies have occurred so recently that their presence or absence is polymorphic when looking at different human populations. The very young AluYa8 subfamily, for example, is estimated to be 50% polymorphic (Roy et al. 1999). The Ya5 lineage is estimated to be around 21% (Otieno et al. 2004), Yg6 around 10% (Salem et al. 2003a), and Yb8 20%, polymorphic (Carroll et al. 2002). Polymorphism data have proven to be useful for human population studies, and have also been used, in conjunction with mutation data for a subfamily, in Alu evolutionary studies. For example, in one case, an Alu element of the Yb8 subfamily has integrated within a polymorphic AluYa5. This is a very recent insertion which has been used to study population structure (Comas et al. 2001). Polymorphic elements, such as Yb8NBC225, have recently been identified as a useful tool for human population studies. There are four alleles at the NBC255 locus, one long form of the Yb8 element, along with two short forms which differ by a single nucleotide polymorphism (SNP), and finally an absence allele (Kass et al. 2007). Although it has been shown that Alus are capable of precise deletion, this is considered to be a rare event, such that absence of an Alu at a particular location would be regarded as an insertion never having occurred.

1.5 CpG Dinucleotides

Alu elements are rich in CpG dinucleotides. 60-90% of these are methylated on cytosine in mammals (Xing et al. 2004), making them prone to spontaneous mutation of the 5'methylcytosine to thymine. This mutation is estimated to occur around six times faster than other types of mutation in Alu elements, ranging from 4.8 to 9.27 times faster in the young subfamilies (Xing et al. 2004). This gives these dinucleotides, and Alu elements in general, a relatively high rate of mutation. Generally, for this reason, when looking at an Alu element subfamily, there will be an excess of TpG and CpA dinucleotides, and a paucity of CpG dinucleotides, relative to the consensus sequence. However, active Alu source genes have retained high numbers of CpGs since the beginning of their proliferation, and it has therefore been postulated that CpGs may play a role in their activity (Jurka et al. 2002). The general Alu consensus sequence contains 21 CpG dinucleotides, but many Alu subfamilies contain many more, for example, the AluYg6 consensus contains

25 CpG dinucleotides. All together, Alu elements contain around 30% of all CpG dinucleotides present in the human genome (Pavlicek et al. 2001).

1.6 Recent research

The identification and study of Alu elements has potential applications in various areas in the study of genome biology. Recent research has focussed on comparative genomics, human disease, exaptation, human populations and broader phylogenetic studies. The investigation and use of Alu elements in so many areas of genome biology makes the availability of complete Alu subfamilies and an understanding of the process of their proliferation and evolution valuable and worthy of further study.

1.6.1 Comparative genomics

Alu elements have been a very popular recent research topic following the release of the complete human and chimpanzee genome sequences. This has enabled whole genome comparative genomic studies to be performed, for example, comparing the distribution of Alus between humans and chimpanzees (Mills et al. 2006). Such studies have shown that the rate of retrotransposition is higher in humans, and that the subfamilies which are most active differ between species. For example, Mills and colleagues found that the AluY and Yc1 subfamilies are most active in chimpanzees, and that the AluYa5 and Yb8 subfamilies, which are highly active in humans, appear inactive. Another study of AluYb8 elements in humans and chimpanzees (Gibbons et al. 2004) revealed only 13 Yb8 elements in chimps, compared with 2201 in humans. The authors suggest a more active reverse transcriptase enzyme in humans as an explanation. The relatively high rate of retrotransposition in humans in relation to chimpanzees has indeed been shown to be due to an increase in rate in humans rather than a decrease along the chimpanzee lineage (Hedges et al. 2004).

1.6.2 Human disease

As there are several Alu source genes that have been shown to be actively retrotransposing, these active elements occasionally give rise to daughters that integrate into human genes, causing disease. It has been estimated that approximately 0.1% of all human genetic disease is due to Alu insertion, and a further 0.3% is due to unequal homologous recombination brought about by Alus (Roy et al. 1999). For example, an Alu retrotransposition-mediated deletion within MEN-1 has been reported, causing multiple endocrine neoplasia type I (Fukuuchi et al. 2006). Inverted Alus appear to be more likely to undergo homologous recombination, and such pairs are consequently found less often than would be expected. One of the few pairs of spatially close inverted Alus that is present in the human genome is frequently deleted in spinal muscular atrophy, due to homologous recombination (Lobachev et al. 2000). Alu elements can also be involved in alternative splicing (see below), and Alport syndrome is caused by aberrant Alu-mediated splicing (Jurka 2004).

1.6.3 Distribution shift

It has been observed for some time that older Alus tend to overrepresented in GC rich (gene rich) regions of the genome, whereas younger elements tend to be found preferentially in AT-rich (gene poor) regions. Several hypotheses have been put forward to explain this observation (Abrusan and Krambeck 2006b;Medstrand et al. 2002). Alu elements insert into the target site TT|AAAA (Jurka and Klonowski 1996), which is the cleavage site of the L1

ORF2p endonuclease. Such sequences are generally found in GC-poor DNA, therefore Alus preferentially insert into GC-poor regions. It is suggested that this preference has been consistent throughout Alu evolution, and that the distribution shift towards increasing GC-content with increasing age of the subfamily can be explained by their effects on the surrounding DNA. Alu elements are capable of recombining with each other, therefore leading to potentially large genomic deletions. Such deletions would be greater tolerated in GC-poor DNA, as these regions tend also to be gene poor (Brookfield 2001). This model is supported by the distribution of Alu elements on the sex chromosomes (Abrusan and Krambeck 2006b). The Y chromosome, for example, which along most of its length can no longer undergo recombination with the X chromosome, shows the original unchanged Alu distribution, with a preference for AT-rich DNA. It is also suggested that the accumulation of Alu elements in GC-rich DNA might reflect paternally-driven selection against elements (Jurka et al. 2004).

L1 elements are found preferentially in AT-rich regions. L1 elements are substantially longer than Alus, and it has been suggested that they would have greater disruptive effects upon insertion into genic, GC-rich regions, and are therefore selected against (Gasior et al. 2007). A study of *de novo* L1 insertions found that only a small window surrounding the target site was GC-poor, and that, beyond this, the flanking sequence exhibited a GC content equivalent to the human genome average of 41% (Gasior et al. 2007). This suggests that the shift of L1 distribution towards AT-rich DNA occurs over evolutionary time, as a result of negative selection. Interestingly, it was also found that *de novo* L1 insertions appear to be clustered, which might suggest that some regions are more susceptible to L1 endonuclease, possibly due to higher order chromatin structure. It has also been shown that newly-inserted

Alu elements are not subject to large amounts of negative selection, as the distributions of fixed and polymorphic members of young Alu subfamilies are very similar (Cordaux et al. 2006b). This suggests that the distribution shift occurs after fixation, again supporting the role of unequal recombination is this process.

1.6.4 Functions and influences on genome evolution

Although Alu elements in general can be regarded as non-functional genomic parasites, which are shown to have negative effects on genome stability, several cases have been identified where an Alu element at a particular locus has been recruited to perform a specific function. This process is referred to as "exaptation". Such functions include involvement in alternative splicing, adenosine-to-inosine (A to I) editing, and regulation of translation (Hasler and Strub 2006). For example, Alu elements contain several potential splice sites (Makalowski 2003), suggesting they can be recruited into coding regions by a process termed "exonisation". In fact, all Alu elements found within exons have been shown to be alternatively spliced. Around 90% of A to I substitutions take place within Alus, with a preference for Alus which have a neighbouring inverted Alu (Hasler and Strub 2006). Due to chromatin condensation, DNA methylation on CpGs, and the weakness of the Alu internal promoter, transcription of Alu elements is generally very low (Li and Schmid 2004). However, under stress conditions, such as viral infection, transcription of Alu RNA is upregulated due to the opening of the chromatin structure, and appears to stimulate general translation at the level of initiation (Li and Schmid 2001).

As well as the instances in which Alus have been recruited to perform a function, they have had an impact on genome and cellular evolution in other ways. Some examples of this impact have already been described, for example, homologous recombination, which can lead to segmental duplications. In fact, a high proportion of Alus (29%) are found at the end of segmental duplications (Jurka 2004;Kazazian 2004), suggesting this might be quite a major effect of the presence of Alu elements. In addition, there are many examples of minisatellites derived from the 5' end of Alu elements (Jurka and Gentles 2006). It is suggested that endonucleolytic attack at the 5' end of the Alu sequence may increase the probability of replication slippage, thereby leading to the formation of minisatellites. Furthermore, Alu elements within the 3' UTR appear to be targeted by some microRNAs. It is suggested that this may be a mechanism of clearing aberrant mRNAs, as Alu elements are most likely to be found within improperly-spliced mRNAs with retained introns (Smalheiser and Torvik 2006).

The methylation of cytosine on many of the CpG dinucleotides within Alu elements has effects related to, for example, the regulation of gene expression and control of development (Xing et al. 2004). Also, as the high numbers of CpGs in Alu elements attracts methylation, this can spread to flanking regions and have other effects. This includes the potential disruption of imprinting, and it might be for this reason that Alus, and SINEs in general, are found at very low levels in imprinted regions (Greally 2002).

1.6.5 Inhibition of retrotransposition

The timing of the expansion of a gene cluster called APOBEC3 (A3) in primates corresponds with a significant, and abrupt, general decrease in retrotransposition (Chiu et al. 2006). The expansion also appears to have occurred before the selective pressure caused by the emergence of primate lentiviruses occurred. It is suggested that the A3 genes, which encode intrinsic antiretroviral proteins, may play a role in limiting retrotransposition rates. Recent studies have indicated that the human endogenous antiretroviral proteins APOBEC3A, 3B (Bogerd et al. 2006) and 3G (Chiu et al. 2006) can inhibit retrotransposition of Alu elements. Alu elements also appear to function as natural targets for A3G, an anti-HIV-1 protein, which sequesters Alu RNA in the cytoplasm, preventing its access to the L1 machinery.

1.6.6 Phylogenetic analyses

Retrotransposon insertions have frequently been used to investigate phylogenetic relationships between species. Examples include the resolution of the human, gorilla and chimpanzee trichotomy (Salem et al. 2003b), and in placing the Cetacea within the Artiodactyla (Shimamura et al. 1997). In addition, Alu elements have been used to investigate the evolutionary history of new world primates (Ray and Batzer 2005). Retrotransposons are useful for this purpose due to their unidirectional evolution, in that the ancestral state is known to be the absence of a retrotransposon at a particular locus, such that Alu insertions generally represent homoplasy-free characters. For example, one study found only three parallel insertions in a set of 500 (Roy-Engel et al. 2002). Generally, L1 and LTR elements are favoured for these kinds of analyses in mammals, as they are present in all mammalian species. For example, both L1 and LTR elements have been used in attempts to determine the correct topology of Superorders in the Eutherian phylogeny. Alu elements are restricted to primates, however, SINEs derived from 7SL RNA are found throughout the Euarchontoglires (Supraprimates). Analysis of these

SINEs established the monophyly of the Euarchonta (Scandentia, Primates and Dermoptera) and of Glires (Rodentia and Lagomorpha) (Kriegs et al. 2007).

1.6.7 Models of Alu amplification

Alu elements were long believed to follow a master gene model of amplification (Roy et al. 2000). Under this model, it is assumed that only one Alu element, at a particular genomic locus, is capable of producing daughter elements (Figure 3.3), which are subsequently inactive, although a few may be capable of low levels of retrotransposition. It assumes that there are certain features of the flanking genomic DNA which favours retrotransposition of the master gene. As the master gene, i.e. the Alu element at this active locus, accumulates mutations, the mutations carried by the daughter elements it produces will change through time, giving rise to the different subfamilies. These mutations that occur in the master gene correspond to the subfamily "diagnostic mutations". These are the mutations that determine which subfamily an Alu element is grouped into. If there truly were an Alu master gene, all bifurcations in the phylogeny would occur along a single branch, generating a pectinate tree (Johnson and Brookfield 2006). In rodents, the master gene producing ID elements was believed to have been identified, in the form of the non-coding RNA BC1, however, it has been shown that this is unlikely to be the only source (Johnson and Brookfield 2006). Although the number of ID elements in each species corresponds to differences at the BC1 locus, in the rat, there are new subfamilies of ID which do not correspond to the sequence of BC1.

It has been shown that at equilibrium, even if several elements within a subfamily are capable of functioning as source genes, if there is a reasonable proportion of element inactivations, a phylogeny will be obtained which suggests a master gene (Brookfield and Johnson 2006). The AluYe lineage supports the master gene model (Salem et al. 2005), as do the results obtained in another study of Alu evolution (Roy-Engel et al. 2002). However, it is evident, from the fact that there are several young Alu subfamilies actively retrotransposing, that the master gene model of Alu evolution is inaccurate. It has also been shown that members of the older AluS subfamilies are still retrotransposing at low levels, as a small number appear to be polymorphic in humans (Mills et al. 2006). It has also been shown that three Alu lineages were active at the time of the platyrrhine-catarrhine divergence – the AluY progenitor, AluSc and AluSp (Ray and Batzer 2005).

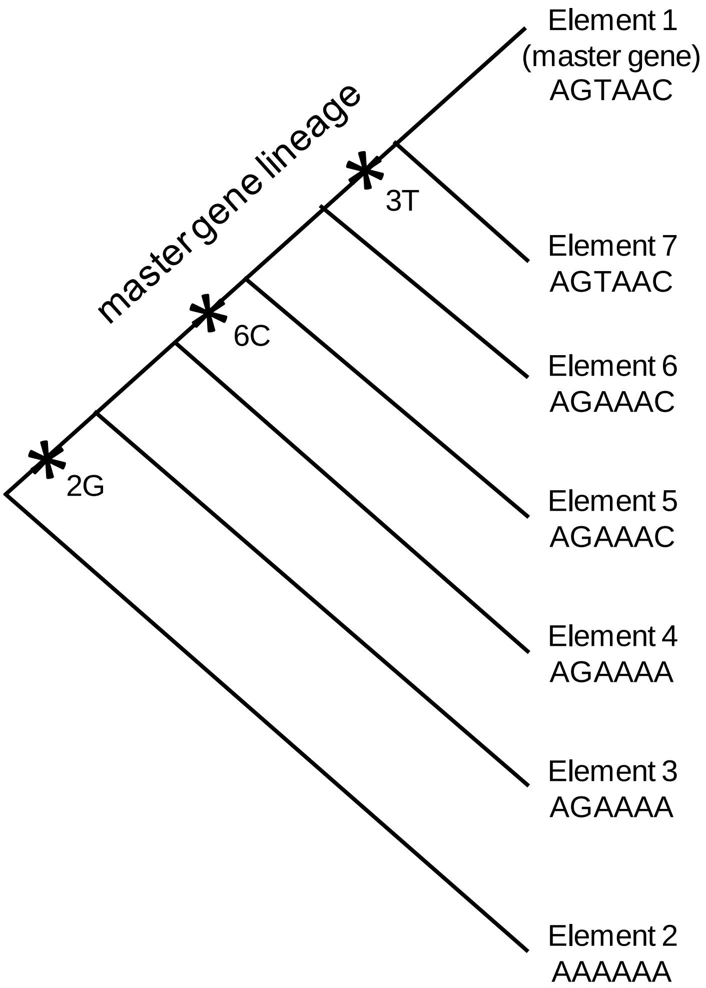


Figure 1.4: Representation of the master gene model of young Alu subfamily expansion. Daughter elements progressively accumulate mutations that occur in the master gene. For simplicity, mutations on the branches leading to daughter elements are not shown.

Several alternatives to the master gene model have been proposed. The most extreme opposite of the master gene model is the “transposon model”, whereby all daughter elements are retrotranspositionally competent. There are also various models which fall somewhere between the two. One such intermediate model of expansion that has been proposed is termed the “stealth model” (Han et al. 2005). This model suggests that an Alu subfamily progenitor arises long before the major expansion of a subfamily, with low-activity source genes (“stealth drivers”) maintaining retroranspositional capability over long periods of time. The source genes would then generate daughter elements with much higher retrotranspositional capability, leading to proliferation of the subfamily.

Another recent study attempted to model Alu subfamily expansion by combining the information from the insertion polymorphism level (IPL), and the nucleotide diversity (π) (Hedges et al. 2005). The IPL corresponds to the proportion of elements within a subfamily which are polymorphic for presence or absence. The authors used computer simulations to determine what level of nucleotide diversity and IPL would be expected under different evolutionary scenarios, ranging from amplification occurring for one million years and then stopping, to occurring for 6 million years, i.e. since the human-chimpanzee divergence. The simulations followed a master gene model. For each alternative model, there was a set of IPL and π parameters that were mutually exclusive. This led to a relative small number of possible amplification histories to explain the real data. Overall, the results rejected the idea of a burst of retrotranspositional activity shortly after the human-chimpanzee divergence.

Assuming that the application of the master gene model is inappropriate for Alu evolution, it is interesting to investigate how many elements within a subfamily are capable of operating as source genes. One such study used the program NETWORK to investigate young Alu subfamily evolution (Cordaux et al. 2004). Unlike traditional methods of inferring the relationships between sequences, NETWORK allows for persistent ancestral nodes and multifurcations, therefore is more appropriate for investigating the relationships between Alu elements, where source genes persist and may give rise to many daughters, than other methods, such as maximum likelihood. The network (Figure 1.5) is produced using a parsimony approach, whereby the relationships inferred are those which required the fewest number of evolutionary changes. This study revealed the existence of a primary node in the network, which corresponds to the source gene giving rise to the majority of subfamily members, and in addition to this, secondary source nodes. The analysis concluded that Alu subfamilies consist of, on average, around 15% secondary source genes, which have given rise to around 30% of the subfamily members. This model of Alu subfamily expansion is referred to as the “sprout” model. A study of Alu RNA levels revealed that only around 100 of the total Alu elements seemed to be producing transcripts (Li and Schmid 2001). Price and colleagues suggest, based on sequence data, that there are 143 Alu source elements in total, which are active at varying rates (Price et al. 2004).

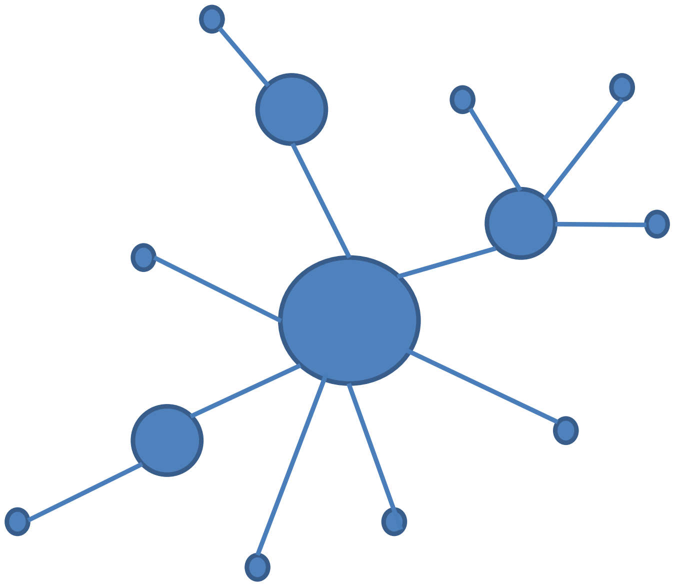


Figure 1.5: An example network of Alu elements. The size of the circles is proportional to the number of Alu elements each circle represents. The large node in the centre is the primary node, which contains the majority of subfamily members. Three relatively large secondary source nodes are observed.

1. 7 Transposable elements in *Drosophila*

The recent availability of the complete genome sequences of twelve species of the *Drosophila* genus makes this group ideal for the investigation of transposable element evolution. The twelve species of *Drosophila* fall into two subgenera, with three species representing the Drosophila subgenus (*D. virilis*, D.*mojavensis* and *D. grimshawi*), and nine species representing the Sophophora subgenus, including the model organism *D. melanogaster* and its relatives *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. persimilis* and *D. willistoni* (a phylogeny of these species is shown in Figure 4.4). These species represent numerous subgroups of the *Drosophila* genus, including the *melanogaster*, *obscura* and *repleta* groups, and share a deep divergence between the two subgenera, between 40 and 60 million years ago. These species span a wide geographical area, with species from the Americas, Asia, Africa and the Pacific Islands, as well as cosmopolitan species, *D. melanogaster* and *D. simulans*, which have recently colonised worldwide (Clark et al. 2007). The quality of the genomic sequence data available varies between species. Species such as *D. melanogaster*, *D. simulans* and *D. virilis* have well-assembled genomes, whereas others are composed of numerous overlapping contiguous sequences (contigs).

The proportion of each genome covered by transposable elements varies between species. Out of the twelve species of *Drosophila* for which the complete sequenced genome is available, transposable element content is highest in *D. ananassae* and *D. willistoni*, at around 25%, but is low in *D. simulans* and *D. grimshawi*, at around only 3% (Clark et al. 2007). 15-18% of the genome of *D. melanogaster* is composed of transposable elements (Vieira and Biemont 2004). This increase in transposable element content in *D.*

melanogaster compared with its close relative *D. simulans* was shown to be due to lower copy number of individual families in *D. simulans*, rather than the absence of entire families from this species. In some cases, such as the LTR retrotransposons Gypsy and Zam, copy numbers were equivalent in both species, which was interpreted to suggest a specific regulatory pathway for these types of transposable element (Vieira and Biemont 2004). Several hypotheses to explain the variation in transposable element copy number between *Drosophila* species have been proposed. The first, which in light of the availability of the twelve *Drosophila* genome sequences can almost certainly be disregarded, is that higher copy numbers of transposable elements are observed in *D. melanogaster* due to the more extensive investigation of this species. Secondly, variations in effective population size are suggested to be involved, such that species with a large effective population size, such as *D. simulans* (Aquadro et al. 1988;Martin-Campos et al. 1992), are more capable of eliminating transposable elements from their genomes. As the selection pressure imposed by individual transposable element copies is small, species with a small effective population size are less able to eliminate elements from their genomes. Thirdly, the *D. simulans* genome is suggested to be more resistant to the proliferation of certain transposable elements, as injection of P elements into *D. simulans* results in a smaller increase in P element copy number than injection into *D. melanogaster* (Kimura and Kidwell 1994). Alternatively, the difference in transposable element copy number between *D. melanogaster* and *D. simulans* is argued to be due to their geographical distribution. As *D. melanogaster* has already colonised globally, but *D. simulans* is still in the process of worldwide colonisation, this may account for the differences in transposable element frequency between them, as *D. simulans* may be exposed to new environments and undergo crosses which might increase transposable element mobilisation (Vieira and Biemont 2004). The *Drosophila* 12 Genomes Consortium determined that the variation in transposable element composition among the twelve genomes correlates significantly with the proportion of each genome composed of euchromatin. Additionally, the highest proportion of the genome covered by transposable elements was observed in species which have the highest number of pseudo-tRNA genes, perhaps suggesting a similar link to that observed in the mouse genome (Waterston et al. 2002). The lowest level of transposable elements was observed in *D. grimshawi*, whose distribution is restricted to Hawaii. It is argued that this species might rarely act as a recipient of a horizontally transferred transposable element, due to its relative isolation compared to the other species, which might consequently result in a lower level of transposable elements in this genome. Transposable element distribution is not continuous throughout *Drosophila* genomes, with regions of both very high and very low transposable element density observed (Bergman et al. 2006). For example, transposable elements are more abundant in pericentromeric regions (Bergman et al. 2006).

1.8 Horizontal transfer

Horizontal transfer is the movement of genetic material from a donor species to a recipient species. This process is very common in prokaryotic systems, but is much rarer in eukaryotes, where the majority of transmission of genetic material is vertical, i.e. from one generation to the next in a single species through the germline. However, there are many documented cases of horizontal transfer involving eukaryotic species, a striking example being the transposable element Space Invader (SPIN) (Gilbert et al. 2009). SPIN has been introduced into the genomes of distantly-related species such as the

African clawed frog, anole lizard, bushbaby and opossum (Gilbert et al. 2009), which are not species typically associated with horizontal transfer of transposable elements.

In *Drosophila*, due to the rapid elimination of transposable elements from the genome, it has been suggested that horizontal transfer may be an essential part of the lifecycle of transposable elements in these species (Bartolome et al. 2009;Jordan et al. 1999;Loreto et al. 2008;Vidal et al. 2009). Contrary to the situation in humans, where transposable elements for the most part represent neutral residents of the genome, in *Drosophila*, transposable elements are strongly selected against and tend not to become fixed in the population. As a consequence, transposable elements in *Drosophila* undergo horizontal transfer to infect a naive host, which has not previously been exposed to that particular transposable element family, and therefore may not have mechanisms in place to limit proliferation of the transposable element. This is followed by a rapid increase in number of the transposable element family, until transposition repression mechanisms evolve to control the element, in many cases leading to the eventual extinction of the family in a particular genome (Silva and Kidwell 2000). A recent study revealed that more than 70% of transposable element families investigated, with representatives from each of the three main types, may have undergone horizontal transfer between only three species: *D. melanogaster*, *D. simulans* and *D. yakuba* (Sanchez-Gracia et al. 2005). However, it has also been argued that horizontal transfer, although likely to have occurred for several transposable element families, is unlikely to have occurred for the majority of families, which are ancient components of many genomes (Lerat et al. 2003). These authors note a high level of sequence similarity between transposable elements, but attribute this to recent increases in transposition rate and high transposable element turnover.

There are several prerequisites to the inference of horizontal transfer of transposable elements. These will be discussed in detail in chapter 4, and include geographical overlap between the putative donor and recipient species, and a vector for transmission (reviewed by Loreto et al. 2008). Horizontal transfer is generally investigated using several lines of evidence. Where the sequence identity between transposable elements of the same family in different species is greater than that observed between host genes, which are known to be under selective constraint, this supports the hypothesis of horizontal transfer. Further evidence can also be considered, such as phylogenetic incongruence, whereby phylogenies constructed using transposable element sequences do not follow the same topology as that known to apply to the host species. This would suggest that the relationships between the transposable element sequences are not the same as the host species. Under vertical transmission, the topologies should be identical. Patchy distribution across the host phylogeny may also be indicative of horizontal transfer, if, for example, a transposable element family is present in a group of species, and a distant relative of those species, but not other species closely related to the putative recipient. Traditionally, particularly in prokaryotes, differences in codon bias have been used to determine horizontal transfer of genetic material from one species to another. However, codon usage is very similar across the twelve *Drosophila* species for which the sequenced genome is available (Clark et al. 2007). Examination of codon usage to infer horizontal transfer also assumes that the host species and the transposable element have the same codon usage, but such a relationship does not appear to exist (Lerat et al. 2000).

Confounding factors, such as selection, stochastic loss and ancestral polymorphism, may lead to similar observations to those expected under horizontal transfer. Selection on transposable elements may lead to effective constraint, such that the sequences do not diverge as much as expected over evolutionary time. Extinction of families in certain lineages, which is a frequent occurrence due to the rapid elimination of elements from *Drosophila* genomes, may also generate a patchy distribution across the phylogeny. Ancestral polymorphism, whereby a variable population of transposable elements is present in a common ancestor, followed by independent assortment, may lead to an incongruent phylogeny in the absence of horizontal transfer. Retention of ancestral polymorphism has been reported, for example, for P elements (Garcia-Planells et al. 1998). The evidence for horizontal transfer, and the alternative explanations for each of the observations which support horizontal transfer, will be discussed in more detail in chapter 4.

1.8.1 Horizontal transfer of DNA transposons

There are many reported cases of horizontal transfer of transposable elements among the *Drosophila*, as well as between *Drosophila* species and other invertebrates (reviewed by Loreto et al. 2008). Perhaps the most well-reported case is of the P element, which was first reported in *D. melanogaster*, where it is associated with hybrid dysgenesis (Bingham et al. 1982;Kidwell et al. 1977), but was found to be absent from closely-related species (Brookfield et al. 1984). The source of the P element in *D. melanogaster*, which is absent from some strains even in this species, was found to be horizontal transfer from *D. willistoni* (Daniels et al. 1990). P elements fall into sixteen subfamilies, patchily distributed across the *Drosophila* genus (Hagemann et al. 1996). For example, two of these sixteen subfamilies, the M and O types, are believed to have undergone horizontal transfer into the *obscura* group. The O type is believed to have transferred into members of the *Drosophila* genus *D. bisfasciata* and *D. imaii* from the *Scaptomyza* (Hagemann et al. 1996). Furthermore a transfer of the O type P element into the ancestor of the *saltans-willistoni* group of the *Drosophila* is supported, although the donor is unknown (Haring et al. 2000), followed by further transfer between species belonging to the *saltans* and *willistoni* groups (de Setta et al. 2007). A recent transfer of the O-type element from the *willistoni* group to the ancestor of the *affinis* subgroup of *Drosophila* is also supported (de Setta et al. 2007). The canonical P element has also been shown to have been involved in horizontal transfer, for example between the *saltans* and *willistoni* groups (Silva and Kidwell 2000). These authors suggest eleven independent horizontal transfer events involving the canonical P element have occurred between these two groups. Transfer of the canonical P element into the ancestor of the *saltans-willistoni* group is also supported (Garcia-Planells et al. 1998). Multiple transfers of the P element among these species is suggested by the presence of multiple subfamilies of P in the *saltans*, *willistoni* and *obscura* groups of the *Drosophila*, although this could be attributed to ancestral polymorphism (Clark et al. 1998). Further cases of horizontal transfer of the P element have been reported (Clark and Kidwell 1997), making it perhaps the most well-documented example of a transposable element family which has undergone horizontal transfer in *Drosophila*. In fact, only a single subfamily of P, the T type, results in a phylogeny which is congruent with host relationships (Haring et al. 1998). The numerous horizontal transfer events, along with ancestral polymorphisms, are likely to account for the range of P element subfamilies distributed across the *Drosophila* genus.

The S element, also a DNA transposon, is present only in *D. melanogaster*, and is reported to have potentially been introduced into this species from an unknown donor, although stochastic loss from the close relatives of *D. melanogaster* is also a possibility (Maside et al. 2003). In this case, the family was not detected in any of nineteen other *Drosophila* species investigated, and therefore any transfer may have occurred from outside the *Drosophila* genus.

Mariner, a widely-distributed family of transposable elements found across many phyla, is also represented in the *Drosophila*, in which it appears to have undergone horizontal transfer. Broadly, Mariner elements from different genera within the Drosophilidae often show greater similarity than mariner elements found in different species of the same genus (Maruyama and Hartl 1991). Horizontal transfer events involving Mariner have been reported to have occurred between *D. mauritiana* and *Z. tuberculatus*, as well as between *Scaptomyza pallida* and members of the *obscura* group of *Drosophila* (Hagemann et al. 1996). Mariner has an unusual distribution in the *Drosophila* genus, in that it is present in *D. mauritiana*, *D. simulans*, *D. sechellia* and *D. yakuba*, but is absent from closely-related species *D. melanogaster*, *D. orena* and *D. erecta*. Phylogenetic incongruence suggests that this unusual distribution may be attributed to horizontal transfer into both the ancestors of *D. simulans* and *D. yakuba* (Brunet et al. 1999).

Minos, a DNA transposon related to Mariner, which is distributed throughout the *Drosophila* genus, but follows a patchy distribution in the Sophophora, also appears to have undergone multiple horizontal transfer events. Transfers between D. *saltans* and the ancestor of *D. mulleri* and *D. mojavensis*, as well as between this ancestor and *D. hydei*, among others, are all supported by sequence similarity and phylogenetic incongruence (de Almeida and Carareto 2005). The patchy distribution of Minos across the Sophophora also supports horizontal transfer from the *repleta* group (Arca and Savakis 2000), with these authors also finding evidence for transfer of Minos from a member of the *repleta* group into the *saltans* group.

1.8.2 Horizontal transfer of LTR retrotransposons

Horizontal transfer was previously thought to be rare among the LTR retrotransposons, with evidence for the process less conclusive than that for DNA transposons (Jordan et al. 1999). However, many cases of horizontal transfer involving LTR retrotransposons have now been confidently reported, and horizontal transfer of LTR retrotransposons is now believed to be more common than for DNA transposons (Bartolome et al. 2009). A famous example of horizontal transfer of an LTR retrotransposon in *Drosophila* is of the Gypsy family and its relatives. It has been suggested that the ability of Gypsy to horizontally transfer may be attributed to the possession of an *env* gene. For example, horizontal transfer of Gypsy virus-like particles produced using the *env* gene between different cells in culture has been demonstrated (Kim et al. 1994;Syomin et al. 2001). However, other types of transposable element, such as the P element, which appear to undergo high rates of horizontal transfer, do not possess the *env* gene. Therefore, it can be assumed to be inessential in this process, although may provide some advantage, as will be discussed in chapter 5. Gypsy appears to have undergone frequent horizontal transfer (Terzian et al. 2000), for example between *D. subobscura* and *D. busckii* (Heredia et al. 2004), which belong to different subgenera of the *Drosophila*. This study postulated a total of nine cases of transmission of Gypsy, ranging from 3.4 to 1.2 million years ago. A close relative of Gypsy, the Gtwin family, appears to have horizontally transferred between *D. melanogaster* and *D. erecta*, with sequence identity of gtwin elements between the two species as high as 99% (Kotnova et al. 2007). These elements appear among others which share more extensive divergence, and appear to have been transmitted vertically. Horizontal transfer of Gtwin from *D. melanogaster* to *D. teissieri* is also supported (Kotnova et al. 2007).

A further famous example of horizontal transfer of an LTR retrotransposon in *Drosophila* is that of the Copia family (Jordan et al. 1999). In this study, Copia elements from a population of *D. willistoni* were found to be more than 99% identical to elements from *D. melanogaster*, strongly supporting horizontal transfer involving these species. Further reported cases include Tirant, an LTR retrotransposon which, although seeming to follow a primarily vertical mode of transmission, appears to have been horizontally transferred from the ancestor of *D. melanogaster* into *D. teissieri*. This event was determined through examination of sequence similarity and phylogenetic incongruence. Mdg3, an LTR retrotransposon which, unlike Gypsy, does not encode *env*, has been shown to undergo horizontal transfer between different cells in culture, and replicate inside the new host cell (Syomin et al. 2002). Other LTR retrotransposons tested (Mdg1, 17.6, 297, 412 and Roo) did not transfer successfully between cells in culture.

1.9 Comparing transposable elements in humans and *Drosophila*

The most striking observation in examining the transposable element content of the genomes of humans and the other Great Apes, compared with the species of the *Drosophila* genus, is the radical difference in copy number. There are, for example, millions of retrotransposons present in humans, compared with only thousands in *Drosophila* (Eickbush and Furano 2002). However, despite a much lower copy number, in *Drosophila*, there are numerous active families of retrotransposons, compared with only one, the L1 family, in humans, which has resulted in the propagation of numerous non-autonomous families such as Alu as well (Eickbush and Furano 2002). Therefore, the types of transposable element found in humans and *Drosophila* also differ considerably. L1 elements comprise 17% of the human genome (Bannert and Kurth 2004), more than the total proportion of the *Drosophila melanogaster* genome comprised of transposable elements of all types. Alu elements comprise around 11% of the human genome, but there are no SINEs found in the *Drosophila melanogaster* genome (Fablet et al. 2007). In total, around 44% of the human genome is comprised of transposable elements (Lander et al. 2001), compared with only around 5% of the euchromatic portion of (Quesneville et al. 2005), or 15-18% of the entire, *Drosophila melanogaster* genome (Vieira and Biemont 2004). The diversity of transposable elements in this genome, currently estimated at sixty families, is greater than that observed in any mammal genome investigated so far (Kapitonov and Jurka 2003a).

The turnover of transposable elements also differs greatly between humans and *Drosophila*. In humans, despite the high copy number of elements, transpositional activity is generally very low. In contrast, transposition is a relatively frequent occurrence in *Drosophila* (Fablet et al. 2007), accounting for around half of observable mutations. The complete human genome sequence is more representative of the entire population in terms of its transposable element content compared with *Drosophila*, where the genome sequence gives an idea of the transposable elements present in a single individual in each species, but variation between individuals is enormous, to the extent that orthologous elements are unlikely to be identified between closely-related species such as *D. simulans* and *D. melanogaster*. For example, only 12.7% of non-LTR retrotransposons in *D. melanogaster* appear to have diverged a significant amount of time since the divergence of *D. melanogaster* and *D. simulans*, and LTR retrotransposons were found to be even more recent (Bergman and Bensasson 2007). Conversely, orthologous elements are commonplace when comparing the genomes of humans and chimpanzees, which possess a similar divergence time to *D. simulans* and *D. melanogaster*. This may, in part, be a consequence of the relatively low rate of ectopic recombination between transposable elements in humans (Eickbush and Furano 2002), among other factors such as effective population size and breeding systems (Fablet et al. 2007). In humans, many transposable element insertions are ancient, whereas, in *Drosophila*, individual transposable element insertions have been present for less than 20 million years (Kapitonov and Jurka 2003a). Regulation of transposable elements also differs between humans and *Drosophila*, with the majority of human regulation occurring at the transcriptional level, for example through epigenetic silencing. Repression of transposition through RNA silencing mechanisms appears to be involved in *Drosophila* (Aravin et al. 2001), as well as nesting of transposable elements producing co-suppression systems (Bergman et al. 2006). Furthermore, in *Drosophila*, selection against new transposable element integrations keeps copy number low (Fablet et al.

2007), which is not a strong influence on the transposable element composition of the human genome. The reduction of inverted Alu repeats compared with expectations does, however, suggest that selection does play a role in determining transposable element composition in humans.

As a consequence of the general retention of transposable elements in humans, inferences can be more easily made regarding the evolutionary history of individual elements. The presence of polymorphic elements in humans, which are present only in some individuals and not others, is also worthy of investigation and provides useful clues into the evolutionary relationships between both transposable elements and different human populations. However, in *Drosophila*, the study of polymorphism is not so informative, as the vast majority of transposable elements are polymorphic within a population (Charlesworth and Langley 1989). For example, in *D. subobscura*, significant differences in the insertion frequencies of the transposable elements Gypsy and Bilbo were found between original and colonising populations, due to founder effects (Garcia Guerreiro et al. 2008). As a consequence of the rapid elimination of transposable elements from *Drosophila* genomes, horizontal transfer is a frequent occurrence for transposable elements in these species, a process which is uncommon, but not unheard of (Gilbert et al. 2009), in primates.

1.10 Aims

The overall aim of this project is to investigate the evolution of transposable elements in the two contrasting systems: humans and *Drosophila*. As discussed previously, the dynamics of transposable element evolution are radically different in these two systems, due to the general retention of elements in humans, compared with the rapid turnover of elements in *Drosophila*. Therefore, the availability of genomic sequence data for these species provides the opportunity to investigate the diverse evolutionary processes affecting transposable elements under different conditions, particularly different levels of selection against the activity of transposable elements.

In humans, and the closely-related species for which genomic sequence data are available, Alu elements, as an example of a type of transposable element in the human genomic environment, were chosen for investigation. The aim of this part of the project was to develop a new method for investigating the evolution of young Alu subfamilies, which have been proliferating in recent history. These sequences were chosen as their recent transposition provides a further source of information for investigating their evolution not available for older sequences, that is, the observation of sequences polymorphic for presence or absence. In addition, these sequences, as recent inhabitants of the genome, have not decayed to the extent of older elements, and therefore can confidently be assigned to particular subfamilies. Traditional phylogenetic reconstruction methods are inappropriate for determining the relationships between Alu elements, as obtaining an accurate phylogeny is difficult, due to the minimal variation between elements in young Alu subfamilies, and frequent parallel mutation, particularly at CpG dinucleotides, which introduces homoplasy. Phylogenetic reconstruction is further complicated by complete and partial gene conversion events, as well as uncertainty regarding the number of source genes and the frequency of complete deletion events. Some previous models of Alu element evolution have assumed an equilibrium situation, whereby old elements are lost at the same rate as the formation of new elements by transposition, resulting in a constant copy number. It is clear that this does not accurately reflect the real situation, and Alu subfamily expansion was one of the aspects of their evolution under investigation. In order to investigate the evolution of young Alu subfamilies, complete sets of sequences of various subfamilies were obtained. The complete sets of sequences belonging to young Alu subfamilies, i.e. subfamilies derived from AluY, were used to investigate the evolution of such subfamilies using computer simulations. Statistics such as pi, theta, and the number of mutations shared between different elements in the same subfamily were used to make inferences about the evolutionary history of each subfamily, such as the number of source elements and the rate at which elements retrotranspose.

To investigate the evolution of transposable elements residing in the genomes of *Drosophila* species, such an approach would be invalid. It is not possible to obtain complete sets of sequences belonging to a particular transposable element family due to the rapid turnover of elements, the rate of which varies between different species. As a consequence of this rapid turnover, transposable elements in *Drosophila* are under much greater selection pressure to avoid elimination. It has been argued that, due to the strength of selection on elements, horizontal transfer of elements has become an "essential part of [their] lifecycle" (Loreto et al. 2008). It appears that for many transposable element families in *Drosophila*, transfer to other, naive, genomes, which have not before encountered a particular family and therefore have not evolved mechanisms to control its proliferation, is a frequent process undertaken to ensure continued survival. As a consequence, horizontal transfer of elements between different species has a major impact on the composition of contemporary transposable element families in *Drosophila* species. The aim of the second part of the project (chapters 4 and

5) is to investigate the process of horizontal transfer of transposable elements in *Drosophila*. This was made possible by the availability of the complete genome sequences of twelve members of the *Drosophila* genus. As copy numbers of transposable elements are much lower in *Drosophila* species compared with humans, rather than investigating a single type of transposable element, as in the case of Alu elements, it is possible to investigate a much wider range of transposable elements in *Drosophila*. Therefore the frequency of horizontal transfer was investigated for the three main types of transposable elements: the LTR and non-LTR retrotransposons (class I) and the DNA transposons (class II).

Chapter 2 – Analysis of the source gene composition and gene conversion in young Alu subfamilies

2.1 Introduction

Alu elements are a family of SINE retrotransposons found in primates, which have been propagated non-autonomously by utilising the enzymatic machinery of autonomous L1 LINE elements (Boeke 1997;Dewannieux et al. 2003). Alu elements are approximately 300bp in length, and have proliferated by the process of retrotransposition (Rogers 1985) to over one million copies (Lander et al. 2001) in the human genome, comprising approximately 11% of the genome by mass (Batzer and Deininger 2002). The majority of these elements were generated 35-60mya during the peak of Alu retrotranspositional activity (Batzer and Deininger 2002), which has subsequently reduced to the current, relatively low level. Despite their high copy number, only a relatively small number of Alu elements are capable of generating new copies (Deininger et al. 1992). This has led to the generation of a collection of Alu subfamilies of differing ages, characterised by diagnostic mutations (Jurka and Milosavljevic 1991). These correspond to mutations present within the source genes that gave rise to each subfamily. The term "source gene" is used to describe an Alu element which is both transcriptionally and retrotranspositionally active, and therefore capable of producing daughter elements. It is known that there are several currently active Alu source genes, each of which has given rise to a "young" Alu subfamily. A subfamily is a collection of Alu elements that have derived from a single source gene, or other active elements descended from that source gene, and therefore share the diagnostic mutations that were present in that source. Several of these young subfamilies, such as AluYg6 (Salem et al. 2003a) and AluYh7, have arisen so recently that subfamily members have only been identified in the genomes of humans, and not of non-human primates.

For a long time, Alu elements were believed to follow a master gene pattern of expansion (Batzer and Deininger 2002), whereby only one, or very few, elements are retrotranspositionally competent. However, although this model appears to be true for some lineages, such as AluYe (Salem et al. 2005), it cannot be true for all of the Alus due to the presence of many currently active source genes, each of which has given rise to a “young” Alu subfamily. For example, it has been reported previously that approximately 10-20% of elements within a young Alu subfamily may operate as secondary source genes (Cordaux et al. 2004). It has also been estimated that there may be at least 143 Alu source genes in total, which would require many active elements within each of the currently-defined subfamilies (Price et al. 2004).

Alu elements, probably as a consequence of their high copy number, undergo relatively frequent gene conversion events (Kass et al. 1995). Gene conversion is a non-reciprocal recombination process, whereby one sequence is converted such that it is identical to a highly similar template sequence, which itself remains unchanged. Gene conversion events involving Alu elements can be complete, whereby the entire element is converted, or partial, such that only a short stretch of sequence within the element is affected.

Following the release of the finalised human genome assembly, it is now possible to obtain, by *in silico* methods, complete sets of Alu sequences belonging to each subfamily. Here, the AluYg, AluYh and AluYi lineages are investigated. A complete set of elements is obtained for each subfamily from the human, and where applicable, chimpanzee, genome, and novel Alu subfamilies are identified. The source gene composition and the influence of gene conversion on the mutational substructure of these subfamilies is also investigated. The activity of a source gene is suggested by the presence of groups of elements with shared combinations of mutations, particularly those groups with elements demonstrating presence/absence polymorphism. The presence of polymorphic elements sharing specific mutations is indicative of the activity of a secondary source element, as polymorphic elements have recently retrotransposed and are therefore unlikely to have accumulated such mutations in parallel. AluYi6 is described as an example of a subfamily which appears to possess numerous secondary source elements, and a novel subfamily, AluYh3a3, is presented as a subfamily which appears to have followed the master gene model of expansion. In light of the data presented in this chapter, the issue of defining what constitutes an Alu subfamily is addressed, along with the criteria that should be followed in assigning an element to a particular subfamily.

In this chapter, the AluYg6, AluYh7, and AluYi6 subfamilies are described in detail, in addition to two newly identified subfamilies, designated AluYh3a1 and AluYh3a3. Firstly, the AluYg6 subfamily is discussed in detail, including putative gene conversion events and the identification of at least two previously unreported secondary source genes. This is followed by discussion of the three families falling on the AluYh lineage, starting with AluYh7, a small subfamily previously reported as AluYh9. The formation of AluYh3a1 and AluYh7 from a single AluYh3 intermediate is discussed, followed by detailed description of the AluYh3a1 subfamily in both humans and chimpanzees. This is followed by discussion of another small subfamily, AluYh3a3, which appears to have derived from AluYh3a1 and followed a master gene model of proliferation. The final subfamily to be discussed is AluYi6, a relatively large subfamily in humans and chimpanzees which appears to possess multiple secondary source elements. Finally, the conventions for assigning Alu elements to particular subfamilies are discussed, including the point at which a secondary source gene of an existing subfamily is considered to be the primary source gene for a novel subfamily.

2.2 Methods

A Basic Local Alignment Search Tool (BLASTN) (Altschul et al. 1997) search, using default parameters, was conducted using the consensus sequences of AluYh9, AluYi6 and AluYg6. A query sequence corresponding to the region between the first and last diagnostic positions was chosen to exclude superfluous sequence from the 5’ and 3’ ends of the consensus which would have increased the number of hits corresponding to Alu elements belonging to other subfamilies. This is particularly true for AluYh3a1, where diagnostic positions are clustered. In addition, the first 47bp of the Alu consensus sequence are identical in the consensus sequences of all but three of the very youngest Alu subfamilies – Yd3, Yd3a1 and Yi6, which contain a C to T transition at position 23 (Jurka et al. 2005). Use of these query sequences, rather than full-length consensus sequences, also reduced the chance of missing genuine members of each subfamily with a substantial 5’ truncation. The search using the AluYh9 consensus revealed only two elements with all nine diagnostic mutations, with the majority of elements sharing only seven. This subfamily will be referred to as AluYh7. Many Alu elements were identified using the AluYh7 consensus sequence, which shared three of the diagnostic mutations of the subfamily, and an additional point mutation. These elements will be referred to as AluYh3a1. This sequence was then used as a query for a BLASTN search.

Each result was examined to check for the presence of subfamily diagnostic mutations. Results were discarded which did not possess the correct base at these diagnostic positions. Results which possessed the correct base at four or greater diagnostic mutations for all but AluYh3a1, in which case two or greater, were retained to investigate the possibility of partial gene conversion events. 1000bp of flanking DNA was extracted 5’ and 3’ of each element, to enable the identification of orthologous regions in closely-related species. These were screened for duplicates, and any sequences which appeared more than once in the dataset were discarded.

Each of the extracted elements and their flanking sequences were submitted as queries for a Blast-like Alignment Tool (BLAT) (Kent 2002) analysis of a related genome (chimpanzee or human) to elucidate complete gene conversion events. Orthologous regions were aligned using ClustalW (Chenna et al. 2003) with default parameters. In most cases, a gap was present in one species corresponding to the region in which the Alu, and one copy of the target site duplication (TSD), were found in the other species. In some cases, an Alu element of an older subfamily was present at the orthologous position, indicative of a gene conversion event. In the case of AluYg6 and AluYh7, which are only found in the human genome, identification of an older Alu element at the orthologous position in the chimpanzee genome is indicative of a gene conversion event along the human lineage. These instances therefore do not represent true AluYg6 and AluYh7 insertions.

Alu elements for all subfamilies were extracted, excluding the poly(A) tail. Each element was given a unique designation consisting of a two-letter code, which is unique to a particular subfamily, followed by a number. The two letter code is used to differentiate between referring to, for example, element 47, and position 47 within an element, following the convention of Salem et al. (2003a). A custom-made Perl program (Appendix 1) was used to identify any mutations that had occurred in each element relative to its subfamily consensus. The program compares a query sequence to a consensus sequence, both provided by input from the user. The output is a list of any mutations that have occurred in the query relative to the consensus, e.g. “A to C transversion at position 44”. Prior to analysis using this program, the two sequences must be aligned. The program recognises dashes (-) in the consensus sequence as insertions and dashes in the query as deletions.

The user input of the query and consensus are stored as two strings, which are then exploded into an array. Consequently, each individual character in the string (each nucleotide in the sequence) is now a single element in an array. The corresponding nucleotides in the consensus and query sequences will occupy the same position in their relative arrays, and therefore each element in the consensus array is compared to the equivalent element in the query array. If the two are identical, the numerical value “identity”, which is initially set to zero, is increased by one. If the nucleotides are different, a specific subroutine is called depending on what type of change has occurred. For example, if an A to C transversion has occurred, the subroutine &AC is called. The program will then print that an A to C transversion has occurred, and the position affected by the mutation.

The NCBI human and chimpanzee genome trace archives were used to assess whether or not each element is polymorphic for presence or absence. In order to do this, for each Alu element for which polymorphism was investigated, the 1000bp of 5' and 3' flanking DNA were joined together, with one copy of the target site duplication, with the Alu element absent. This sequence of approximately 2000bp was then used as a query for a search of the trace archives, which contain the unassembled genome sequences of the individuals which were sequenced in order to produce the final human and chimpanzee genome assemblies. As the archives contain sequences corresponding to various individuals, it is possible that if an Alu insertion is polymorphic, the sequence of individuals in which the insertion is absent may be present. Therefore, if a sequence is found that matches the query, but is not interrupted by an Alu insertion, this insertion can be inferred to be polymorphic for presence or absence. As only a limited number of individuals are represented in the trace archives, particularly in the case of the chimpanzee, this will result in an underestimation of polymorphism, as some polymorphic elements that are present at high frequency in the population may not be detected.

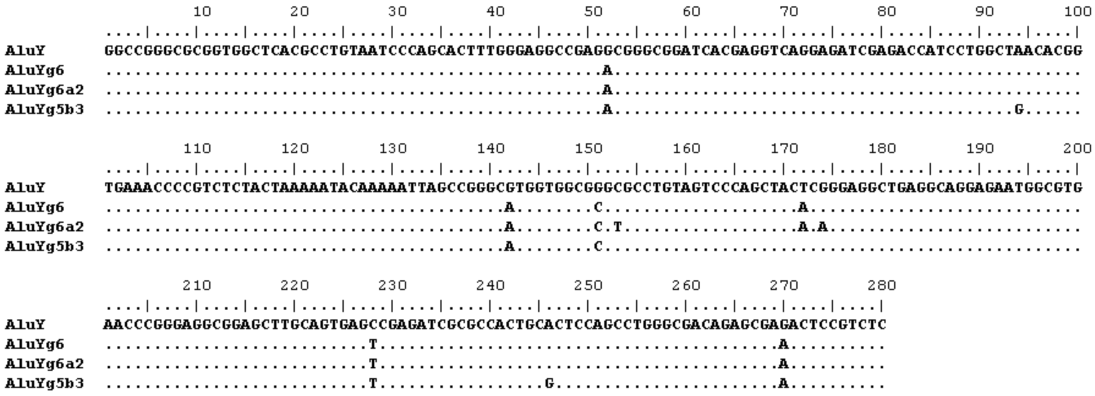
2.3 Results and Discussion

The number of source elements contributing to young Alu subfamilies appears to vary widely among lineages. The AluYg6 subfamily appears to possess several source elements, with varying levels of activity. The AluYh lineage appears to have split into two, which share three diagnostic mutations in addition to those of AluY. There is evidence for master gene expansion within the AluYh lineage, as mutations appear to have accumulated progressively in one subfamily. This provides further evidence that the master gene model remains consistent with the pattern of proliferation in some young Alu subfamilies. The AluYi lineage provides an alternative perspective, with evidence for multiple secondary source elements simultaneously contributing to the proliferation of the AluYi6 subfamily. This is suggested by the presence of multiple elements within a subfamily sharing a set of specific mutations, which suggests these mutations were present in the source element rather than occurring multiple times in parallel. The hypothesis is supported by the presence of elements polymorphic for presence or absence, which share additional mutations from the AluYi6 consensus. This suggests that these mutations were present in the source gene which gave rise to these elements, rather than happening multiple times in parallel. However, using the genome trace archives to identify polymorphisms cannot conclusively determine whether or not an element which appears to be fixed is polymorphic, as individuals in which the element is absent may not be represented in the archives. It is therefore possible that the number of polymorphic elements, and therefore potentially the number of secondary source elements, has been underestimated. Gene conversion appears to have influenced the structure of mutations observed in all three lineages investigated, in one case, resulting in the inactivation of a putative master gene. These results are discussed in detail for the AluYg, AluYh and AluYi lineages below.

2.3.1 AluYg6

The AluYg6 subfamily consensus is 281bp in length, and is characterised by six diagnostic changes from the AluY consensus (Figure 2.1). A total of 380 AluYg6 elements were extracted from the human genome (Styles and Brookfield 2007). 281 of these possessed all six AluYg6 diagnostic mutations, including 23 that matched the AluYg6 consensus perfectly. In addition to these 281, a further 11 elements exhibited 5 of the diagnostic mutations, with a non-ancestral change at the final position, which can be assumed to have been generated by a forward mutation event at the diagnostic position. This generates a set of 292 Alu elements which have unequivocally been derived from a source gene of the AluYg6 subfamily.

The other 88 elements show an ancestral base, in other words that found in the AluY consensus, at one or two of the diagnostic positions. Such elements may have been generated by back mutation, gene conversion, or more likely a mixture of these two processes. Of these 88 sequences, 71 showed an ancestral base at only one of the six diagnostic sites. In 29 cases, this single diagnostic change was the presence of an ancestral T at position 172, however, this large number can be explained by the inference of a new source gene carrying this mutation (see below). Only four of these 29 sequences do not appear to have been derived from this source gene, and are included in table 2.1.



```
                    10        20        30        40        50        60        70        80        90       100
            ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
AluY        GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTAACACGG
AluYg6      ...................................................A................................................
AluYg6a2    ...................................................A................................................
AluYg5b3    ...................................................A.........................................G......

                   110       120       130       140       150       160       170       180       190       200
            ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
AluY        TGAAACCCCGTCTCTACTAAAAATACAAAAATTAGCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTG
AluYg6      .........................................A........C....................A............................
AluYg6a2    .........................................A........C.T..................A.A..........................
AluYg5b3    .........................................A........C.................................................

                   210       220       230       240       250       260       270       280
            ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
AluY        AACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTC
AluYg6      ...........................T.........................................A..........
AluYg6a2    ...........................T.........................................A..........
AluYg5b3    ...........................T.................G.......................A..........
```

Figure 2.1: Alignment of the consensus sequences of the AluY, AluYg6, AluYg6a2 and AluYg5b3 subfamilies. Diagnostic mutations are shown for the three younger subfamilies. Identical nucleotides are represented by dots.

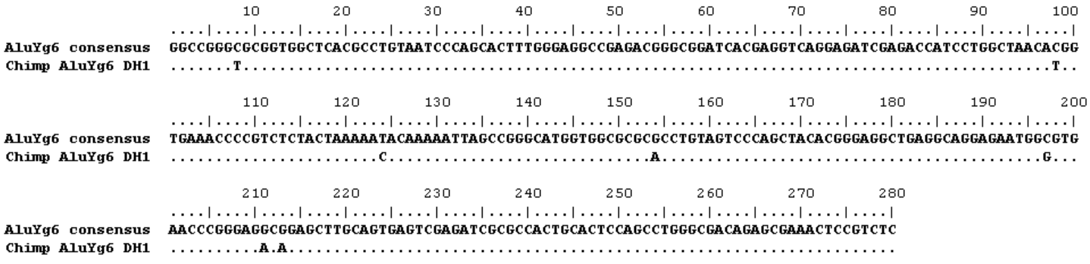
It was found that ancestral bases were overrepresented compared to expectations at diagnostic positions in elements with one or two diagnostic changes relative to other non-Yg6 bases, i.e. where the nucleotide at a diagnostic position was not that found in the AluYg6 consensus, this was more likely to be the nucleotide observed in the AluY consensus, rather than either of the other two nucleotides (chi-squared test, $p < 0.05$). This suggests at least some instances of partial gene conversion, whereby short gene conversion tracts have modified part of an AluYg6 insertion using an older Alu element as a template. It was also found that the ancestral bases that could be generated from the AluYg6 diagnostic bases by transition mutations were more common than those that could be generated by transversions. Taken together, these two observations suggest that both processes, back mutation and partial gene conversion, have each in some way contributed to the diversity seen among elements of the AluYg6 subfamily.

Position	**Ancestral base**	**Yg6 base**	**Nature of back mutation**	**Occurrence**
52	G	A	Transition	8
142	G	A	Transition	10
151	G	C	Transversion	4
172	T	A	Transversion	4
228	C	T	Transition	7
270	G	A	Transition	13

Table 2.1: Frequency of ancestral bases in AluYg6 elements with one diagnostic change.

The ancestral base 270G occurs most frequently of the six ancestral bases in these single diagnostic position variants. This might suggest that mutation of the ancestral G to an A at this position was the final mutation to occur along the AluYg lineage, and that some of these sequences represent intermediate “AluYg5” elements. However, if elements with two diagnostic changes are also considered, 142G is the most common ancestral base, occurring 20 times in total, with 270G being the second most common (15 times in total). Both 52G and 142G occur frequently in single position variants, and the lower frequency of both 151T and 172T can be explained by the relative unlikelihood of back mutations at these positions, as these would require transversional rather than transitional changes. The fact that there is not much difference in the frequency of the ancestral bases among AluYg6 elements with one diagnostic change might indicate an absence of intermediates, suggesting that the AluYg lineage did not become retrotranspositionally active until all six diagnostic changes had occurred.

As expected for a young Alu subfamily (Pavlicek et al. 2001), AluYg6 elements appear to integrate preferentially into AT-rich DNA. 5’ truncations are relatively common in AluYg6 elements, brought about by incomplete reverse transcription or by imprecise integration (Salem et al. 2003a). They do not represent post-integration deletion events. 35 AluYg6 elements were truncated at the 5’ end, with truncations ranging from 5 to 67bp. The mean length of these truncations is 30bp, and the modal length, exhibited by 5 elements, is 36bp.



```
                          10        20        30        40        50        60        70        80        90       100
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
AluYg6 consensus  GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGACGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTAACACGG
Chimp AluYg6 DH1  .......T.........................................................................................T..

                         110       120       130       140       150       160       170       180       190       200
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
AluYg6 consensus  TGAAACCCCGTCTCTACTAAAAATACAAAAATTAGCCGGGCATGGTGGCGCGCGCCTGTAGTCCCAGCTACACGGGAGGCTGAGGCAGGAGAATGGCGTG
Chimp AluYg6 DH1  .......................C.............................A..........................................G...

                         210       220       230       240       250       260       270       280
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
AluYg6 consensus  AACCCGGGAGGCGGAGCTTGCAGTGAGTCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAAACTCCGTCTC
Chimp AluYg6 DH1  ..........A.A...................................................................
```

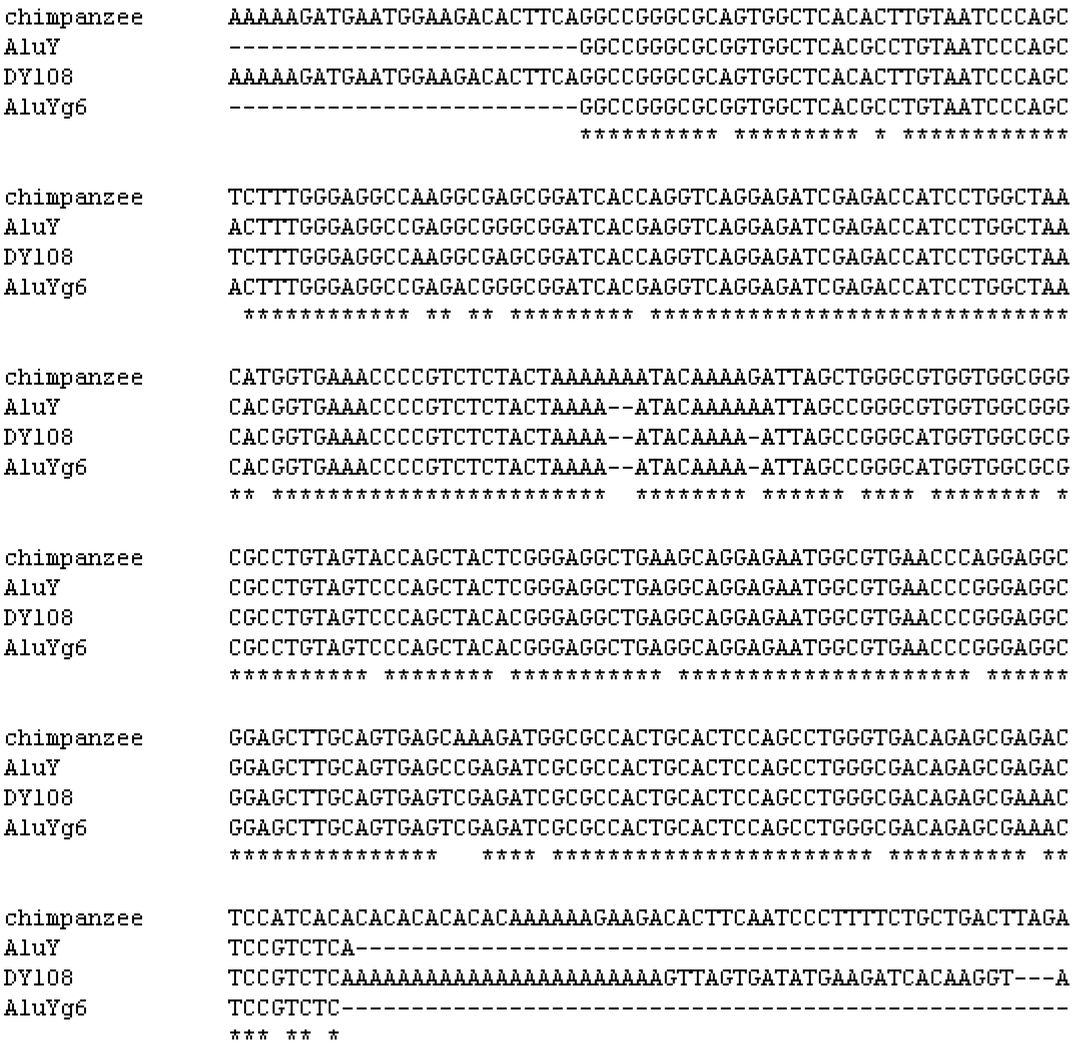
Figure 2.2: Alignment of the chimpanzee AluYg6 sequence (DH1) and the human AluYg6 consensus.

In contrast to expectations, as AluYg6 has been reported to be human-specific, one example of an AluYg6 element was identified in the chimpanzee. This element possesses all six AluYg6 diagnostic mutations, along with seven additional mutations (Figure 2.2). Only one of these seven is found in the consensus sequence of another Alu subfamily on Repbase Update. This is the 124T to C mutation, which is found in the AluYf2 consensus. Four of the additional mutations are CpG transitions. It is therefore much more likely to have integrated as an AluYg6 element than one belonging to any other subfamily. If it had not integrated as AluYg6, the six diagnostic mutations would have had to occur by chance, which is unlikely as these are not particularly common mutations, for example, none are CpG transitions. This AluYg6 is not present at the orthologous region in the human genome, where there is only one copy of the target site duplication. This element may have transposed to its current location in the chimpanzee following the human-chimpanzee divergence, which would indicate that there is at least one other AluYg6 in the chimpanzee. Alternatively, it is possible that this Alu was present in the human-chimpanzee ancestor and has been precisely deleted by recombination between the flanking direct repeats along the human lineage, a property of Alu elements that has been identified before (van de Lagemaat et al. 2005). Alternatively, this AluYg6 may have been polymorphic in the ancestral population, and has been fixed in the chimpanzee but lost by drift in humans. Regardless of which of these explanations is correct, this finding shows that the first AluYg6 element must have arisen further into the past than previously estimated (Salem et al. 2003a), although the subfamily may have undergone a period of relative dormancy with respect to its retrotranspositional rate (Han et al. 2005), only proliferating to considerable numbers along the human lineage following the human-chimpanzee divergence. It has previously been suggested that the evolution of a successful subfamily progenitor sequence occurs well in advance of its peak activity (Hedges et al. 2004). These authors also note the lower levels of young Alu insertions in the chimpanzee relative to the human genome, and suggest a general increase in retrotranspositional activity in humans as the most favourable explanation.

To look for evidence of partial gene conversion events, the frequency of putative back mutations at the six diagnostic positions (for elements with one or two diagnostic changes) was compared to the frequency of the other possible mutations at these sites. Changes to the ancestral AluY base are greatly overrepresented relative to the alternative two bases at each position, except in the case of 151C, which shows the ancestral G in 8 cases, and a non-ancestral T in 15 cases. However, this site is within a CpG dinucleotide, which explains why a T is seen so frequently at this position. CpG transition mutations occur at approximately six times the rate of non-CpG mutations (Xing et al. 2004) due to spontaneous deamination of 5-methylcytosine to thymine, resulting in a paucity of CpG, and an excess of TpG and CpA dinucleotides, as in this case. It is also noteworthy that although the transversional change to the ancestral nucleotide is seen 8 times, the alternative transversion is not seen at all. Perhaps the best evidence supporting the occurrence of partial gene conversion events is at position 172. Excluding elements carrying the ancestral mutation believed to have arisen in another source gene, as above, 8 back mutations are seen, which would represent transversional changes. There were no instances of the other transversion (A to C) seen at this position, and only four instances of the transition mutation.

For nine of the elements identified, an Alu was present at the orthologous locus in the chimpanzee, indicative of a complete gene conversion event. This is where an Alu element belonging to an older subfamily has been converted to an AluYg6 along the human lineage. As well as the three complete gene conversion events previously reported (Salem et al. 2003a), six more were identified. In five of these cases (DY108, DY178, DY198, DY285 and DY364), a complete Alu element is present in the chimpanzee (Figure 2.3).

The final case, DY184, is more ambiguous, as only the left monomer and a short section of the right monomer of an AluSx element are present in the chimpanzee. The human AluYg6 sequence is flanked at the 3' end by a 17bp region of homology to the part of the right monomer that remains in the chimpanzee. It is likely that homologous recombination has occurred between these two 17bp regions along the chimpanzee lineage, causing most of the 3' end of the AluSx to be deleted, and leaving only one copy of the homologous region.



```
chimpanzee      AAAAAGATGAATGGAAGACACTTCAGGCCGGGCGCAGTGGCTCACACTTGTAATCCCAGC
AluY            -------------------------GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGC
DY108           AAAAAGATGAATGGAAGACACTTCAGGCCGGGCGCAGTGGCTCACACTTGTAATCCCAGC
AluYg6          -------------------------GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGC
                                         ********** ********* * ************

chimpanzee      TCTTTGGGAGGCCAAGGCGAGCGGATCACCAGGTCAGGAGATCGAGACCATCCTGGCTAA
AluY            ACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTAA
DY108           TCTTTGGGAGGCCAAGGCGAGCGGATCACCAGGTCAGGAGATCGAGACCATCCTGGCTAA
AluYg6          ACTTTGGGAGGCCGAGACGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTAA
                 ************ ** ** ********* ******************************

chimpanzee      CATGGTGAAACCCCGTCTCTACTAAAAAAATACAAAAGATTAGCTGGGCGTGGTGGCGGG
AluY            CACGGTGAAACCCCGTCTCTACTAAAA--ATACAAAAAATTAGCCGGGCGTGGTGGCGGG
DY108           CACGGTGAAACCCCGTCTCTACTAAAA--ATACAAAA-ATTAGCCGGGCATGGTGGCGCG
AluYg6          CACGGTGAAACCCCGTCTCTACTAAAA--ATACAAAA-ATTAGCCGGGCATGGTGGCGCG
                ** ************************  ******** ****** **** ******** *

chimpanzee      CGCCTGTAGTACCAGCTACTCGGGAGGCTGAAGCAGGAGAATGGCGTGAACCCAGGAGGC
AluY            CGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGC
DY108           CGCCTGTAGTCCCAGCTACACGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGC
AluYg6          CGCCTGTAGTCCCAGCTACACGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGC
                ********** ******** *********** ********************* ******

chimpanzee      GGAGCTTGCAGTGAGCAAAGATGGCGCCACTGCACTCCAGCCTGGGTGACAGAGCGAGAC
AluY            GGAGCTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGAC
DY108           GGAGCTTGCAGTGAGTCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAAAC
AluYg6          GGAGCTTGCAGTGAGTCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAAAC
                ***************   **** *********************** ********** **

chimpanzee      TCCATCACACACACACACACAAAAAAGAAGACACTTCAATCCCTTTTCTGCTGACTTAGA
AluY            TCCGTCTCA---------------------------------------------------
DY108           TCCGTCTCAAAAAAAAAAAAAAAAAAAAAAAAGTTAGTGATATGAAGATCACAAGGT---A
AluYg6          TCCGTCTC----------------------------------------------------
                *** ** *
```

Figure 2.3: Alignment of human DY108 (and flanking sequence) with the orthologous region from the chimpanzee. The consensus sequences for AluY, found in the chimpanzee, and AluYg6 are also shown. Alignments were performed with ClustalW using default settings, followed by manual editing.

In two cases (DY178 and DY184), the gene conversion event appears to be complete. It is possible that parallel insertion, rather than gene conversion, is responsible for this observation. This is, however, unlikely in the case of element DY184, as the element in chimpanzee belongs to the AluSx subfamily, which is believed to only be currently retrotransposing at extremely low levels (Johanning et al. 2003). In the other four cases, one ancestral base is present in the AluYg6 sequence (52G), suggesting gene conversion tracts of approximately 200bp have converted the majority of the sequence, but the beginning of the element is still ancestral. These four cases can be inferred to represent almost complete gene conversion events, although in the absence of the element in the chimpanzee, these would more likely be interpreted as short gene conversion tracts having converted a stretch of bases at the beginning an AluYg6 using an older Alu as a template. This alternative explanation is still possible, as the element in the chimpanzee may represent a parallel insertion, and the site may have been unfilled in human-chimpanzee ancestor. An AluYg6 would then have inserted along the human lineage, which was then partially converted to AluY. If this were the case, given that all four confirmed partial gene conversion tracts cover the 5’ end of the sequence, it might suggest a preference for gene conversion tracts forming in this region. A preference for gene conversion of the beginning of the element may be explained by a greater degree of homology among Alu elements in this region. Information regarding the nature of this site in the genomes of other African apes would help to resolve this issue.

It is likely that there are examples of complete gene conversion that cannot be detected. For example, such events may have generated AluYg6 elements from other young Alu elements, which would not be present in the chimpanzee. It is also possible that backward gene conversion events have occurred, whereby following its insertion, an AluYg6 has been converted to an older Alu element, and therefore cannot be identified as an AluYg6 insertion.

2.3.1.1 AluYg6a2

40 elements were identified with all six of the AluYg6 diagnostic mutations along with two additional mutations (153T and 174A). Both of these mutations are CpG transitions, and may therefore be expected to be observed frequently without the inference of a new source gene. However, based on the frequency of these mutations occurring independently in the rest of the data, it was found that these two mutations are found together within a single element significantly more often than would be expected due to parallel mutation (chi-squared test, p<0.01), therefore suggesting a source element containing these two mutations is responsible for their propagation. These elements will be referred to as Yg6a2, according to the standard nomenclature for Alu elements (Batzer et al. 1996). An unfilled site was seen at the orthologous locus in the chimpanzee genome for all Yg6a2 elements.

The CpG mutation at position 153 occurs 33 times in the rest of the data (a total of 73 times including Yg6a2 elements), whereas the CpG mutation at position 174 only occurs twice in the rest of the data (42 times in total). This might suggest that the 153C to T mutation occurred first, within the source gene, and was propagated before the 174G to A mutation took place within the same source gene.

2.3.1.2 AluYg5b3

27 elements were identified which have five diagnostic mutations of the AluYg6 subfamily, along with three additional ones. These three mutations occur significantly more frequently together than alone in the complete Yg6 dataset, suggesting these mutations are shared by descent from a new source gene, rather than by multiple parallel mutation events. Interestingly, a

further mutation is shared by seven of the Yg5b3 elements, which is only seen in fifteen of the 353 remaining AluYg6 elements. This may indicate that this mutation has occurred in the Yg5b3 source gene and has been subsequently proliferated. However, the mutation is a C to T transition occurring within a CpG dinucleotide, which may have arisen multiple times independently. An unfilled site was seen at the orthologous locus in the chimpanzee genome for all Yg5b3 elements.

One of the three mutations that are diagnostic for this new subfamily is a back mutation at one of the six AluYg6 diagnostic sites (172A to T). Yg5b3 is therefore an appropriate designation for this subfamily, as it contains five of the diagnostic mutations of the AluYg lineage, along with three mutations which distinguish it from its ancestral sequence, AluYg6. It is also possible that this subfamily may be derived from an intermediate AluYg5, which began retrotransposing following the occurrence of the two additional mutations.

Active Alu source genes generally appear to have retained high numbers of CpG dinucleotides, which will have degenerated to TpG and CpA in inactive elements. As CpG dinucleotides are prone to rapid degeneration, elements with no CpG mutations may represent recent transpositions. The AluYg6 subfamily consensus sequence contains 25 CpG dinucleotides, compared to 26 in AluYg5b3. This high level of CpG may be related to the activity of the source gene.

AluYg6a2 elements generally have a high level of identity to their consensus sequence, with approximately 33% (13/40) showing perfect identity to the consensus. This might suggest a relatively recent origin for this subfamily. In contrast, only around 7% (2/27) of AluYg5b3 elements are identical to their subfamily consensus, which is a similar proportion to the AluYg6 subfamily in general. The fact that there are proportionately fewer elements in the Yg5b3 subfamily that are identical to their consensus might suggest that they have been around for some time and simply retrotranspose relatively inefficiently. It is unlikely that either of these two subfamilies represent mutations occurring in the AluYg6 source gene, as recent retrotranspositions of this gene have been identified in the form of low frequency polymorphic insertions (Salem et al. 2003a).

Although no other source genes within the AluYg6 subfamily have propagated to the extent of the two described above, there are a further two groups of elements for which inference of secondary source genes is a possible explanation for their shared mutations. 23 elements were identified which have a C at position 277, instead of the T found in the Yg6 consensus. This would represent a transitional mutation, and may therefore be expected to occur relatively frequently, but it is highly overrepresented relative to other transition mutations. The second group contains only two elements, but the rarity of the mutations they share makes parallel mutation unlikely. Elements DY380 on chromosome 18 and DY383 on chromosome 19 both show a G to A mutation at position 11, a 3-mer expansion of the middle A-rich tract, and a two nucleotide insertion ("AC") at position 173. Although the other two mutations are relatively common, small insertions into AluYg6 elements, which do not correspond to poly(A) tract expansion, are extremely rare, occurring in only six other elements out of 380. Out of these six cases, only one possesses a dinucleotide insertion. Four possess single nucleotide insertions and the final one an eight nucleotide duplication. In this case, it is quite likely that gene conversion is responsible for the shared variation in these two elements rather than retrotransposition. This may be a more favourable explanation as the mutations are shared by only two elements, although these sequences could suggest a source gene active at very low levels. The high frequency of the 277C mutation is much more likely to represent the activity of another source gene, although alternative explanations are also possible, such as a high rate of mutation at this site. However, if this were the case, the other non-ancestral nucleotides (A and G) would also be expected to be seen at high frequency, and this is not the case. It is possible that this mutation has occurred frequently by chance, and with only one mutation shared between elements it is harder to distinguish between source gene activity and parallel mutation.

2.3.1.3 Polymorphism

AluYg6 was assessed for the presence of polymorphic elements which might indicate the activity of secondary source genes. Polymorphic elements belonging to the derivative subfamilies Yg6a2 and Yg5b3 were identified (Table 2.2), revealing that these subfamilies are at least 30% and 26% polymorphic, respectively. Polymorphic elements matching the AluYg6 consensus sequence were also identified. In addition to polymorphic elements corresponding to the Yg6a2 and Yg5b3 consensus sequences, polymorphic elements were identified with the diagnostic mutations for each of these subfamilies along with an additional shared mutation in each case. Three of the five AluYg6a2 elements, and two out of seven of the AluYg5b3 elements, with the additional mutation, were found to be polymorphic. The presence of polymorphic elements with additional shared mutations suggests that the number of source elements in the AluYg6 subfamily may be at least five, although the consensus sequence of only a single AluYg6 source gene has previously been reported (Salem et al. 2003a).

Subfamily	Species	Mutations from Yg6 consensus	Copy number	Polymorphic?
Yg6	H, C	-	380 (H) 1 (C)	Yes
Yg6a2	H	153T, 174A	40	Yes
Yg5b3	H	94G, 172T, 246G	27	Yes

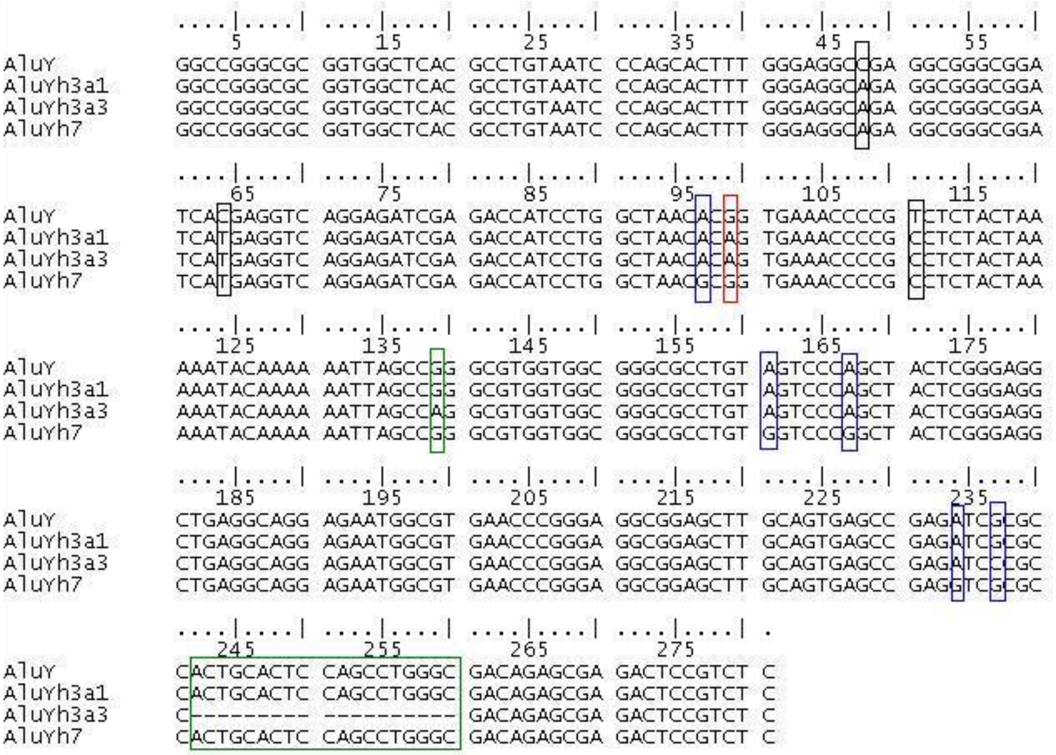
Table 2.2: AluYg6 derivative subfamilies. Copy number of the AluYg6 subfamily includes elements belonging to the two derivative subfamilies. H = human, C = chimp.

2.3.2 The AluYh lineage

2.3.2.1 AluYh7

The elements of the AluYh9 subfamily (Jurka et al. 2002) all share only seven diagnostic mutations, therefore this subfamily will be referred to as AluYh7 (Figure 2.4). The subfamily is human-specific and contains twenty elements (Styles and Brookfield 2009), of which sixteen have been previously reported as AluYh9 (Jurka et al. 2002). This subfamily appears to have arisen very recently, as at least half of the elements are polymorphic for presence or absence, and nine of the elements are identical to the subfamily consensus. The level of divergence of the remaining eleven elements is very low, with elements possessing either one or two point mutations from the consensus. Of the two elements which possess the nine diagnostic mutations of AluYh9, one is polymorphic for presence or absence. This makes it likely that these additional two mutations are shared due to retrotransposition rather than parallel mutation or gene conversion, so there may be two active source genes in this small subfamily.

The only evidence for proliferation on this lineage prior to the acquisition of all seven diagnostic mutations of AluYh7 is of an element with three of the seven diagnostic mutations, “AluYh3”, which appears to have generated two derivative lineages, one of which is AluYh7. The second shares these three diagnostic mutations with AluYh3, along with an additional mutation, and shall be referred to as AluYh3a1 (Figure 2.5).



```
               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   5         15         25         35         45         55
AluY           GGCCGGGCGC GGTGGCTCAC GCCTGTAATC CCAGCACTTT GGGAGGCCGA GGCGGGCGGA
AluYh3a1       GGCCGGGCGC GGTGGCTCAC GCCTGTAATC CCAGCACTTT GGGAGGCAGA GGCGGGCGGA
AluYh3a3       GGCCGGGCGC GGTGGCTCAC GCCTGTAATC CCAGCACTTT GGGAGGCAGA GGCGGGCGGA
AluYh7         GGCCGGGCGC GGTGGCTCAC GCCTGTAATC CCAGCACTTT GGGAGGCAGA GGCGGGCGGA

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   65         75         85         95        105        115
AluY           TCACGAGGTC AGGAGATCGA GACCATCCTG GCTAACACGG TGAAACCCCG TCTCTACTAA
AluYh3a1       TCATGAGGTC AGGAGATCGA GACCATCCTG GCTAACACAG TGAAACCCCG CCTCTACTAA
AluYh3a3       TCATGAGGTC AGGAGATCGA GACCATCCTG GCTAACACAG TGAAACCCCG CCTCTACTAA
AluYh7         TCATGAGGTC AGGAGATCGA GACCATCCTG GCTAACGCGG TGAAACCCCG CCTCTACTAA

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  125        135        145        155        165        175
AluY           AAATACAAAA AATTAGCCGG GCGTGGTGGC GGGCGCCTGT AGTCCCAGCT ACTCGGGAGG
AluYh3a1       AAATACAAAA AATTAGCCGG GCGTGGTGGC GGGCGCCTGT AGTCCCAGCT ACTCGGGAGG
AluYh3a3       AAATACAAAA AATTAGCCAG GCGTGGTGGC GGGCGCCTGT AGTCCCAGCT ACTCGGGAGG
AluYh7         AAATACAAAA AATTAGCCGG GCGTGGTGGC GGGCGCCTGT GGTCCCGGCT ACTCGGGAGG

               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  185        195        205        215        225        235
AluY           CTGAGGCAGG AGAATGGCGT GAACCCGGGA GGCGGAGCTT GCAGTGAGCC GAGATCGCGC
AluYh3a1       CTGAGGCAGG AGAATGGCGT GAACCCGGGA GGCGGAGCTT GCAGTGAGCC GAGATCGCGC
AluYh3a3       CTGAGGCAGG AGAATGGCGT GAACCCGGGA GGCGGAGCTT GCAGTGAGCC GAGATCCCGC
AluYh7         CTGAGGCAGG AGAATGGCGT GAACCCGGGA GGCGGAGCTT GCAGTGAGCC GAGGTCGCGC

               ....|....| ....|....| ....|....| ....|....| .
                  245        255        265        275
AluY           CACTGCACTC CAGCCTGGGC GACAGAGCGA GACTCCGTCT C
AluYh3a1       CACTGCACTC CAGCCTGGGC GACAGAGCGA GACTCCGTCT C
AluYh3a3       C--------- ---------- GACAGAGCGA GACTCCGTCT C
AluYh7         CACTGCACTC CAGCCTGGGC GACAGAGCGA GACTCCGTCT C
```

Figure 2.4: Alignment of the AluYh7, AluYh3a1 and AluYh3a3 consensus sequences with AluY. Diagnostic mutations from AluY can be seen for each subfamily. Mutations from AluY shared by all three subfamilies on the AluYh lineage are shown in black boxes. The further mutations accumulated by the AluYh7 subfamily are shown in blue boxes. The mutation possessed by AluYh3a1 and AluYh3a3 is shown in a red box, and the further two mutations in AluYh3a3 are shown in green boxes.

2.3.2.2 AluYh3a1

It can be assumed that the AluYh3a1 subfamily is derived from the putative “AluYh3” intermediate along this lineage (Figure 2.5). Although it is possible that these three mutations from AluY occurred twice independently, this is more unlikely. All four of the diagnostic mutations for this subfamily are found in the left half of the element (Figure 2.5). AluYh3a1 appears to have originated before the divergence of humans, chimpanzees and gorillas, as there are instances of elements of this subfamily present in the gorilla whole genome shotgun sequence. There are also instances of AluYh3a1 present in the pre-ensembl release of the orangutan genome, where there are at least three elements present. These elements are not found in humans. However, the subfamily appears to be absent from the available genomic data for two species of gibbon (*Hylobates concolor* and *Namascus leucogenys*). If the available gibbon sequence data are representative of the whole genome, this would suggest the AluYh3a1 subfamily originated between around 10 and 16 million years ago.

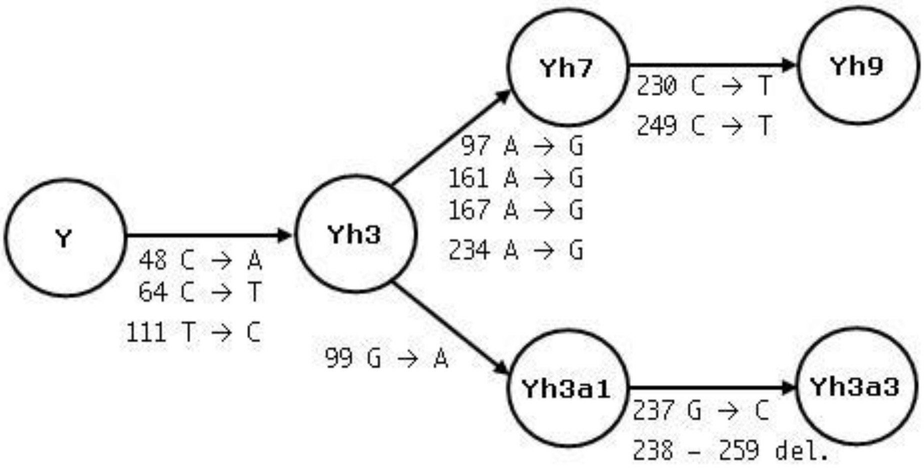


Figure 2.5: Relationships between the subfamilies on the AluYh lineage. Diagnostic mutations for each new subfamily are shown on the arrow leading to that subfamily. The copy numbers of each of these subfamilies are listed in Table 2.3.

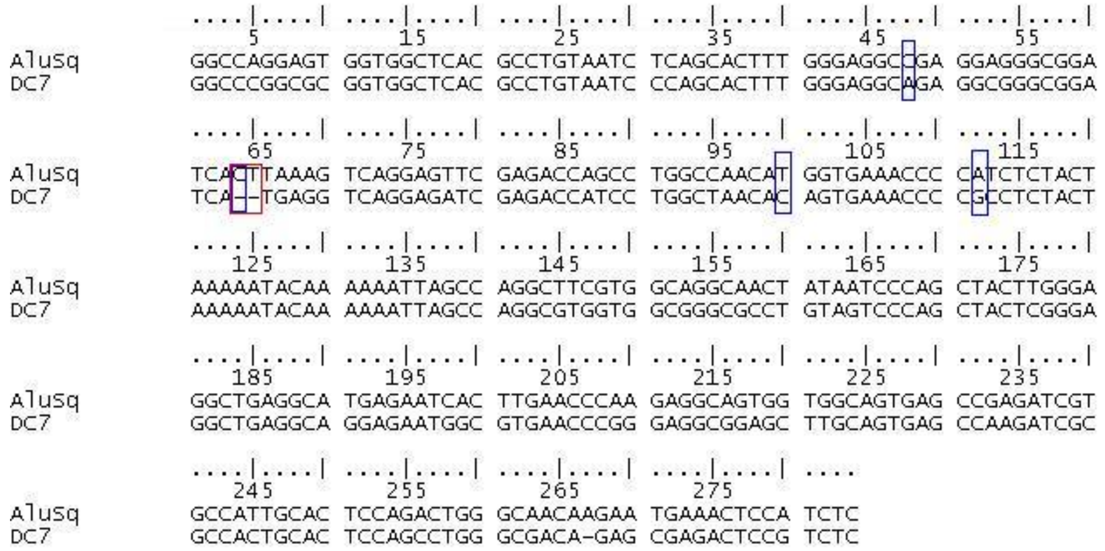
The subfamily has proliferated quite extensively in humans and chimpanzees, and many elements are shared between the two species. There are 98 elements with all four AluYh3a1 diagnostic mutations present in humans, and

73 in chimpanzees (Styles and Brookfield 2009). Out of the 73 elements found in chimpanzees with all four diagnostic mutations, only 16 are unique to chimpanzees, with the remaining 57 found in both chimpanzees and humans. It is not unexpected that there would be more human-specific than chimp-specific elements, as the rate of retrotransposition has been shown to have increased along the human lineage (Mills et al. 2006), with most young Alu subfamilies that are present in both species reaching larger copy numbers in humans than in chimpanzees.

Putative subfamily	Species	Mutations from AluYh3 consensus	Copy number	Polymorphic?
Yh9	H	97G, 161G, 167G, 230T, 234G, 249T	2	Yes
Yh7	H	97G, 161G, 167G, 234G	20	Yes
Yh3a1	H, C, G, O	99A	98 (H), 73 (C)	-
Yh3a3	H, C	99A, 237C, (238-259 del.)	3 (H), 11 (C)	Unknown

Table 2.3: Subfamilies on the AluYh lineage. Copy number of the AluYh7 subfamily includes elements with the diagnostic mutations for AluYh9. H = human, C = chimp, G = gorilla, O = orangutan. Copy number in the gorilla and orangutan genomes is unknown due to the absence of complete genome sequences for these species. Polymorphism of AluYh3a1 was not tested.

A complete gene conversion event has occurred in chimpanzees. There is an AluYh3a1 present in the chimpanzee (DC7), but an older AluSq element is found at the orthologous locus in the human genome (Figure 2.6). This is likely to be a forward gene conversion event in the chimpanzee rather than a backwards event in the human due to the high similarity between DC7 and the shared element DC8/DY83, which is likely to have provided the template. There are other examples of possible complete gene conversion events occurring between species-specific Alu elements, where pairs of elements share numerous mutations. However, as the putative gene conversion events would be occurring between two species-specific elements, it cannot be proven that the mutations are not shared due to parallel mutation.

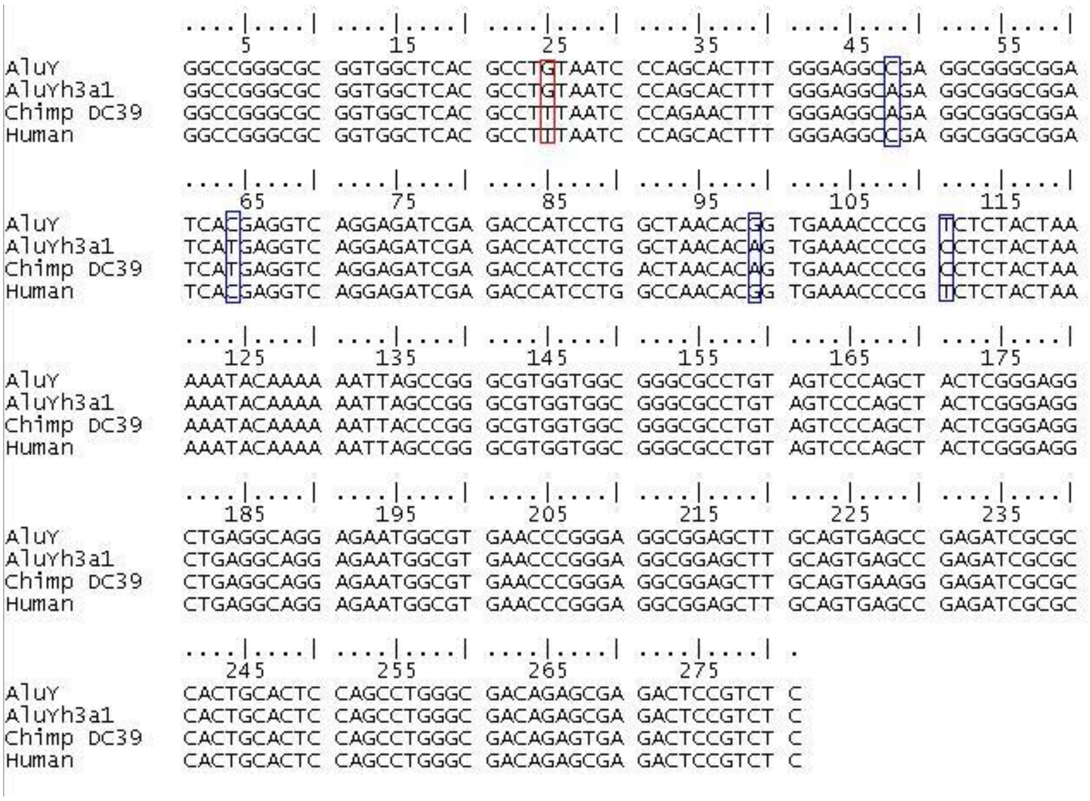


```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    5         15         25         35         45         55
AluSq           GGCCAGGAGT GGTGGCTCAC GCCTGTAATC TCAGCACTTT GGGAGGCCGA GGAGGGCGGA
DC7             GGCCCGGCGC GGTGGCTCAC GCCTGTAATC CCAGCACTTT GGGAGGCAGA GGCGGGCGGA

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                    65         75         85         95        105        115
AluSq           TCACTTAAAG TCAGGAGTTC GAGACCAGCC TGGCCAACAT GGTGAAACCC CATCTCTACT
DC7             TCA--TGAGG TCAGGAGATC GAGACCATCC TGGCTAACAC AGTGAAACCC CGCCTCTACT

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   125        135        145        155        165        175
AluSq           AAAAATACAA AAAATTAGCC AGGCTTCGTG GCAGGCAACT ATAATCCCAG CTACTTGGGA
DC7             AAAAATACAA AAAATTAGCC AGGCGTGGTG GCGGGCGCCT GTAGTCCCAG CTACTCGGGA

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   185        195        205        215        225        235
AluSq           GGCTGAGGCA TGAGAATCAC TTGAACCCAA GAGGCAGTGG TGGCAGTGAG CCGAGATCGT
DC7             GGCTGAGGCA GGAGAATGGC GTGAACCCGG GAGGCGGAGC TTGCAGTGAG CCAAGATCGC

                ....|....| ....|....| ....|....| ....|....| ....
                   245        255        265        275
AluSq           GCCATTGCAC TCCAGACTGG GCAACAAGAA TGAAACTCCA TCTC
DC7             GCCACTGCAC TCCAGCCTGG GCGACA-GAG CGAGACTCCG TCTC
```

Figure 2.6: Alignment of the chimpanzee AluYh3a1 element DC7 and the AluSq element present at the orthologous position in the human genome. Diagnostic positions for AluYh3a1 are shown in blue boxes, the characteristic deletion of AluSq is shown in a red box. This case represents an example of complete gene conversion replacing an Alu element from an old subfamily with one from a younger subfamily.

In addition, patterns of mutations suggest multiple partial and “almost complete” gene conversion events have occurred. Comparison of elements shared by humans and chimpanzees reveals that ancestral nucleotides have been introduced at diagnostic positions in one species. This may be either due to partial gene conversion or back mutation. In two cases, all four diagnostic sites possess the ancestral nucleotide in one species, but this is likely to be due to partial gene conversion, rather than complete, as the orthologues share mutations outside the putative gene conversion tract

(Figure 2.7). In the case of AluYh3a1, the diagnostic mutations are clustered within a 64bp region. It is therefore reasonable that a partial gene conversion tract, which on average cover around 50-100bp (Batzer and Deininger 2002;Roy-Engel et al. 2002), would result in ancestral nucleotides being introduced at all four sites. There are further examples of likely partial gene conversion events resulting in between one and three diagnostic mutations. Once again, mutations are shared between orthologues on either side of the putative tract, but no mutations are shared within it.



```
            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                5          15         25         35         45         55
AluY        GGCCGGGCGC GGTGGCTCAC GCCTGTAATC CCAGCACTTT GGGAGGCCGA GGCGGGCGGA
AluYh3a1    GGCCGGGCGC GGTGGCTCAC GCCTGTAATC CCAGCACTTT GGGAGGCAGA GGCGGGCGGA
Chimp DC39  GGCCGGGCGC GGTGGCTCAC GCCTTTAATC CCAGAACTTT GGGAGGCAGA GGCGGGCGGA
Human       GGCCGGGCGC GGTGGCTCAC GCCTTTAATC CCAGCACTTT GGGAGGCCGA GGCGGGCGGA

            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                65         75         85         95        105        115
AluY        TCACGAGGTC AGGAGATCGA GACCATCCTG GCTAACACGG TGAAACCCCG TCTCTACTAA
AluYh3a1    TCATGAGGTC AGGAGATCGA GACCATCCTG GCTAACACAG TGAAACCCCG CCTCTACTAA
Chimp DC39  TCATGAGGTC AGGAGATCGA GACCATCCTG ACTAACACAG TGAAACCCCG CCTCTACTAA
Human       TCACGAGGTC AGGAGATCGA GACCATCCTG GCCAACACGG TGAAACCCCG TCTCTACTAA

            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
               125        135        145        155        165        175
AluY        AAATACAAAA AATTAGCCGG GCGTGGTGGC GGGCGCCTGT AGTCCCAGCT ACTCGGGAGG
AluYh3a1    AAATACAAAA AATTAGCCGG GCGTGGTGGC GGGCGCCTGT AGTCCCAGCT ACTCGGGAGG
Chimp DC39  AAATACAAAA AATTACCCGG GCGTGGTGGC GGGCGCCTGT AGTCCCAGCT ACTCGGGAGG
Human       AAATACAAAA AATTAGCCGG GCGTGGTGGC GGGCGCCTGT AGTCCCAGCT ACTCGGGAGG

            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
               185        195        205        215        225        235
AluY        CTGAGGCAGG AGAATGGCGT GAACCCGGGA GGCGGAGCTT GCAGTGAGCC GAGATCGCGC
AluYh3a1    CTGAGGCAGG AGAATGGCGT GAACCCGGGA GGCGGAGCTT GCAGTGAGCC GAGATCGCGC
Chimp DC39  CTGAGGCAGG AGAATGGCGT GAACCCGGGA GGCGGAGCTT GCAGTGAAGG GAGATCGCGC
Human       CTGAGGCAGG AGAATGGCGT GAACCCGGGA GGCGGAGCTT GCAGTGAGCC GAGATCGCGC

            ....|....| ....|....| ....|....| ....|....| .
               245        255        265        275
AluY        CACTGCACTC CAGCCTGGGC GACAGAGCGA GACTCCGTCT C
AluYh3a1    CACTGCACTC CAGCCTGGGC GACAGAGCGA GACTCCGTCT C
Chimp DC39  CACTGCACTC CAGCCTGGGC GACAGAGTGA GACTCCGTCT C
Human       CACTGCACTC CAGCCTGGGC GACAGAGCGA GACTCCGTCT C
```

Figure 2.7: Alignment of chimpanzee AluYh3a1 element DC39 and the human orthologue. A partial gene conversion event has introduced diagnostic mutations into the human element, shown in blue boxes. A transversion mutation is shared between orthologues outside the putative gene conversion tract, shown in the red box. It is possible that one of these elements has been introduced by complete gene conversion, with subsequent parallel mutation at position 25.

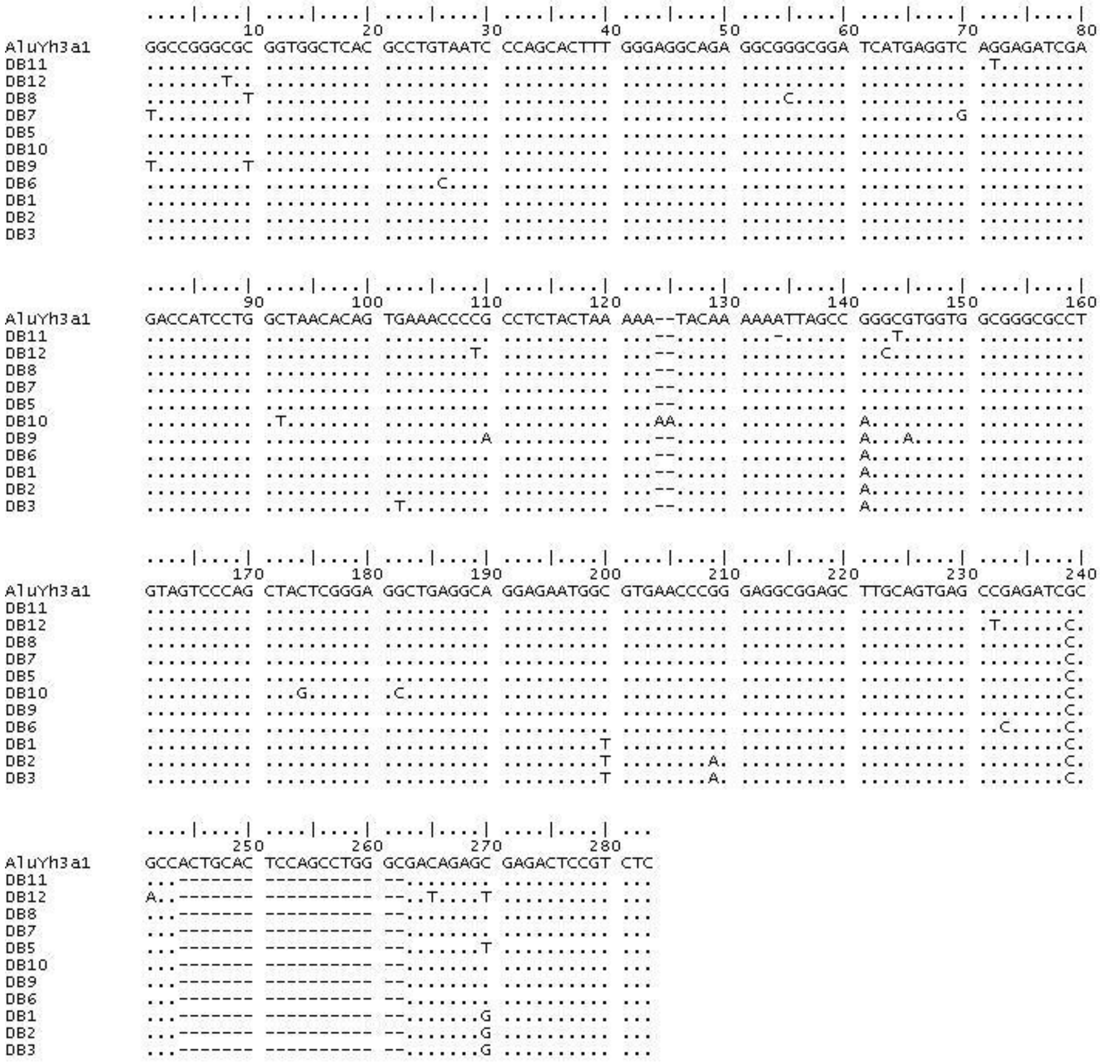
There is no compelling evidence from the sequence data for this subfamily to suggest that secondary source elements have contributed to its proliferation. The greatest number of elements sharing a mutation within a species is 14 of the 98 elements. However, this mutation is a CpG transition, which is likely to have occurred many times in parallel. The 98 sequences in this subfamily differ from the consensus sequence by 2.622 mutations, on average, at non-CpG sites. This represents a difference of 0.01107 mutations per base. Assuming that CpG sites are six times as likely to change as non-CpG bases, then, using the binomial distribution, the probability that a given CpG base in the sequence would be mutated in 14 or more of the sequences is 0.532%. Almost all of these mutations can be assumed to be CpG transitions. There are 44 CpG bases in the sequence, each of which has a chance of 0.532% of being mutated in 14 or more of the sequences. Therefore, the probability that at least one of these 44 would be mutated in 14 or more of the 98 sequences is approximately 44 times this, or more than 20%. The observation that this site mutated in 14 sequences is consistent with a hypothesis of 14 independent mutational events. Given the relatively small size of this subfamily, however, the high incidence of this mutation may be indicative of a secondary source element. The greatest number of elements sharing a pair of mutations is five, which is, again, possibly due to the presence of a secondary source element possessing these two mutations, but could be due to parallel mutation. Levels of polymorphism were not assessed for this subfamily due to time constraints. There was no evidence from the sequence data to suggest the presence of secondary source elements, and therefore without a candidate secondary source sequence, every element in the family would have to have been examined for presence/absence polymorphism. As a consequence of not analysing polymorphism, it is not possible to make any predictions about the number of secondary source elements which may be functional in this subfamily, i.e., it cannot be assumed that a single master gene has produced all members of this family.

2.3.2.3 AluYh3a3

There is a small subfamily which appears to have been derived from AluYh3a1 (Table 2.3), but beyond this, there is no evidence for substructure within the AluYh3a1 subfamily to indicate the activity of further secondary source elements. Therefore, it is possible that the remaining elements in this subfamily have been produced by the activity of a single source (master) gene. Alternatively, there may be several source elements, which do not possess mutations, or perhaps only a single CpG mutation, which alone would not provide enough evidence to suggest the activity of a secondary source gene.

The derivative subfamily, named AluYh3a3, contains a characteristic 19bp deletion near the 3’ end, between positions 242 and 260. This subfamily is very small, comprising eleven elements in chimpanzees and only three elements in humans (Styles and Brookfield 2009). However, large deletions in Alu elements are rare and so the presence of the deletion in these elements, in addition to the four diagnostic mutations of AluYh3a1, is good enough evidence to consider this a unique subfamily, rather than due to parallel deletion. In addition to the 19bp deletion, ten of the eleven chimpanzee elements also contain a diagnostic point mutation. AluYh3a3 is unusual in that it has proliferated to a greater extent in chimpanzees than humans. As was described for AluYh3a1, it is much more common for a subfamily to be more prevalent in humans, as the general rate of retrotransposition has increased along the human lineage (Mills et al. 2006). Two of the elements are shared between chimpanzees and humans.

Although AluYh3a3 consists of a very small number of elements, there is a considerable amount of substructure in this subfamily. The pattern of shared mutations can be explained under the master gene model, without the inference of secondary source elements, as there appears to have been a progressive accumulation of mutations (Figure 2.8, Figure 2.9). The putative master element would be either DB2 or DB3, as these two elements possess all of the shared mutations. However, neither of these two elements is shared between chimpanzees and humans. The orthologous locus in humans is a perfect unfilled site for DB2, showing a single copy of the target site duplication. The orthologous locus in humans to chimp DB3 is a filled site containing an AluSx element. It is therefore possible that DB3 is the master gene for the AluYh3a3 subfamily, but the locus has undergone backwards gene conversion in humans. Gene conversion has been reported to be more likely between spatially close Alu elements (Zhi 2007), and the human DB3 orthologue is found in a region with numerous highly similar AluSx elements (possessing the characteristic 20bp deletion) in the vicinity that could have provided the gene conversion template. Presence of this deletion might have made gene conversion more likely between this AluSx element and AluYh3a3, which contains a 19bp deletion, as the sequences would be of similar length.



```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                        10         20         30         40         50         60         70         80
AluYh3a1        GGCCGGGCGC GGTGGCTCAC GCCTGTAATC CCAGCACTTT GGGAGGCAGA GGCGGGCGGA TCATGAGGTC AGGAGATCGA
DB11            .......... .......... .......... .......... .......... .......... .......... .T........
DB12            .......T.. .......... .......... .......... .......... .......... .......... ..........
DB8             .........T .......... .......... .......... .......... ....C..... .......... ..........
DB7             T......... .......... .......... .......... .......... .......... .........G ..........
DB5             .......... .......... .......... .......... .......... .......... .......... ..........
DB10            .......... .......... .......... .......... .......... .......... .......... ..........
DB9             T........T .......... .......... .......... .......... .......... .......... ..........
DB6             .......... .......... .....C.... .......... .......... .......... .......... ..........
DB1             .......... .......... .......... .......... .......... .......... .......... ..........
DB2             .......... .......... .......... .......... .......... .......... .......... ..........
DB3             .......... .......... .......... .......... .......... .......... .......... ..........

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                        90        100        110        120        130        140        150        160
AluYh3a1        GACCATCCTG GCTAACACAG TGAAACCCCG CCTCTACTAA AAA--TACAA AAAATTAGCC GGGCGTGGTG GCGGGCGCCT
DB11            .......... .......... .......... .......... ...--..... ...-...... ...T...... ..........
DB12            .......... .......... ........T. .......... ...--..... .......... ..C....... ..........
DB8             .......... .......... .......... .......... ...--..... .......... .......... ..........
DB7             .......... .......... .......... .......... ...--..... .......... .......... ..........
DB5             .......... .......... .......... .......... ...--..... .......... .......... ..........
DB10            .......... .T........ .......... .......... ...AA..... .......... A......... ..........
DB9             .......... .......... .........A .......... ...--..... .......... A...A..... ..........
DB6             .......... .......... .......... .......... ...--..... .......... A......... ..........
DB1             .......... .......... .......... .......... ...--..... .......... A......... ..........
DB2             .......... .......... .......... .......... ...--..... .......... A......... ..........
DB3             .......... .......... .T........ .......... ...--..... .......... A......... ..........

                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                       170        180        190        200        210        220        230        240
AluYh3a1        GTAGTCCCAG CTACTCGGGA GGCTGAGGCA GGAGAATGGC GTGAACCCGG GAGGCGGAGC TTGCAGTGAG CCGAGATCGC
DB11            .......... .......... .......... .......... .......... .......... .......... ..........
DB12            .......... .......... .......... .......... .......... .......... .......... .T......C.
DB8             .......... .......... .......... .......... .......... .......... .......... ........C.
DB7             .......... .......... .......... .......... .......... .......... .......... ........C.
DB5             .......... .......... .......... .......... .......... .......... .......... ........C.
DB10            .......... ...G...... .C........ .......... .......... .......... .......... ........C.
DB9             .......... .......... .......... .......... .......... .......... .......... ........C.
DB6             .......... .......... .......... .......... .......... .......... .......... ..C.....C.
DB1             .......... .......... .......... .........T .......... .......... .......... ........C.
DB2             .......... .......... .......... .........T ........A. .......... .......... ........C.
DB3             .......... .......... .......... .........T ........A. .......... .......... ........C.

                ....|....| ....|....| ....|....| ....|....| ...
                       250        260        270        280
AluYh3a1        GCCACTGCAC TCCAGCCTGG GCGACAGAGC GAGACTCCGT CTC
DB11            ...------- ---------- --........ .......... ...
DB12            A..------- ---------- --..T....T .......... ...
DB8             ...------- ---------- --........ .......... ...
DB7             ...------- ---------- --........ .......... ...
DB5             ...------- ---------- --.......T .......... ...
DB10            ...------- ---------- --........ .......... ...
DB9             ...------- ---------- --........ .......... ...
DB6             ...------- ---------- --........ .......... ...
DB1             ...------- ---------- --.......G .......... ...
DB2             ...------- ---------- --.......G .......... ...
DB3             ...------- ---------- --.......G .......... ...
```

Figure 2.8: Alignment of AluYh3a3 elements in the chimpanzee, with the AluYh3a1 consensus. A progressive accumulation of mutations can be seen in the putative master gene, DB3, supporting the master gene model of proliferation for this subfamily.

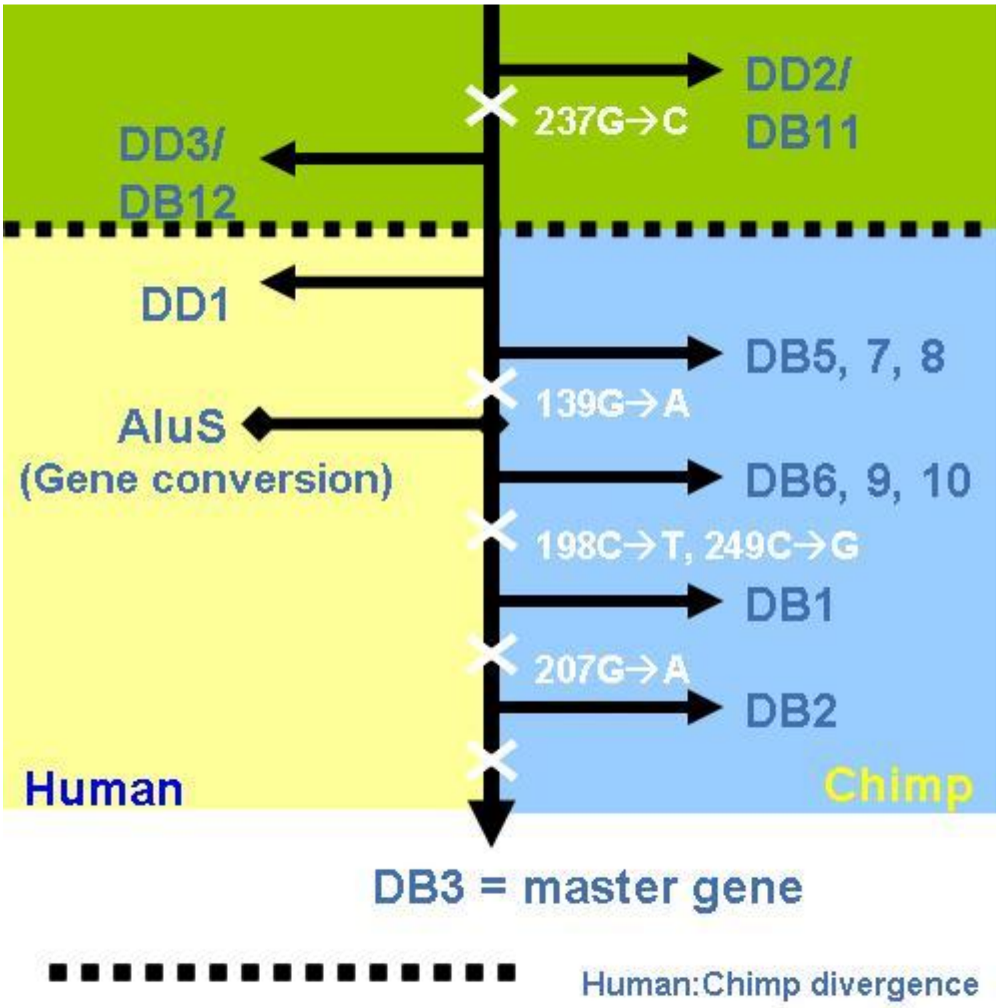


Figure 2.9: Progressive accumulation of mutations in the AluYh3a3 master gene. The master gene model of proliferation accounts for the sharing of mutations by several groups of elements in the AluYh3a3 subfamily in chimpanzees. A complete gene conversion event appears to have inactivated the master gene along the human lineage.

Inactivation of the master gene by gene conversion would also explain why there are fewer elements of this subfamily in humans compared with chimpanzees. This is more likely to be a backwards gene conversion on the human lineage rather than a forward gene conversion in the chimpanzee, as the site is unfilled in the orangutan genome. AluSx currently mobilises at only very low frequency (Johanning et al. 2003), and therefore is more likely to have been introduced to this locus in humans by gene conversion rather than retrotransposition. AluYh3a3 is also present in the gorilla, however, the orthologous locus to chimp DB3 is unavailable to determine whether or not this is the original founder element. There are four copies of AluYh3a3 in the available gorilla genomic data, all of which are gorilla-specific. No AluYh3a3 elements appear to be polymorphic in either humans or chimpanzees by examination of the trace archives. This may indicate that the subfamily is no longer proliferating, or individuals with polymorphic elements may not be represented in the archives. This is more likely for the chimpanzee sequences, as fewer individuals have been sequenced.

2.3.3 AluYi6

The AluYi6 subfamily, which has six diagnostic mutations from the AluY consensus, has been reported to be present in humans, chimpanzees and gorillas (Salem et al. 2003a). 123 elements belonging to this subfamily have been reported, 104 of which possess all six diagnostic mutations (Salem et al. 2003a). In this study, 237 Yi6 elements were identified in humans (Styles and Brookfield 2009). The sequences of these elements show patterns of shared mutations consistent with the activity of potentially many secondary source elements. Three derivative subfamilies, designated Yi6.1, Yi6.2 and Yi6.3 have already been reported, and have all been shown to be polymorphic, and therefore currently active (Salem et al. 2003a). The pattern of shared mutations in the AluYi6 subfamily indicates there may be as many as fourteen source elements operating in humans (Figure 2.10, Table 2.4). These potential fourteen source genes fall on only three lineages, as each possesses one of three mutations: 151T, 57T or 254A. Polymorphism data suggest that this is not indicative of three “master genes”, but does indeed represent the activity of many source elements. Some of these small Yi6 derivative subfamilies contain a considerable number of elements (for example, there are 36 elements with the 57T mutation), whereas others contain very few. The potential derivative subfamilies which only contain very few elements, such as those with 254A and 251T mutations (five elements), may not be the product of secondary source genes, as the mutations may simply be shared due to parallel mutation. Polymorphism data, however, suggest the former is more likely.

It was shown previously that three of the elements found in humans were shared with the chimpanzee (Salem et al. 2003a). Analysis of the chimpanzee genome reveals that Yi6 has proliferated quite extensively in the chimpanzee following its divergence from humans. 91 Yi6 elements were found to be present in the chimpanzee genome, of which thirteen are shared with humans. It is to be expected that there would be fewer copies of the subfamily in chimpanzees relative to humans, due to the general increase in retrotransposition rate along the human lineage. Only three of the chimpanzee elements are identical to the AluYi6 subfamily consensus. Two of the previously identified derivative subfamilies (Yi6.1 and Yi6.2) were found in both humans and chimps, suggesting a time of origin prior to the human-chimp divergence. A novel chimp-specific subfamily was also identified, the consensus for which has two additional diagnostic mutations (175A, 200A) relative to the AluYi6 consensus. 31 of the 91 chimpanzee Yi6 elements belong to this novel subfamily. In chimpanzees, at least three AluYi6 source genes appear to be currently active from looking at polymorphism data, containing the 151T mutation, the 175A and 200A mutations, and the 57T mutation.

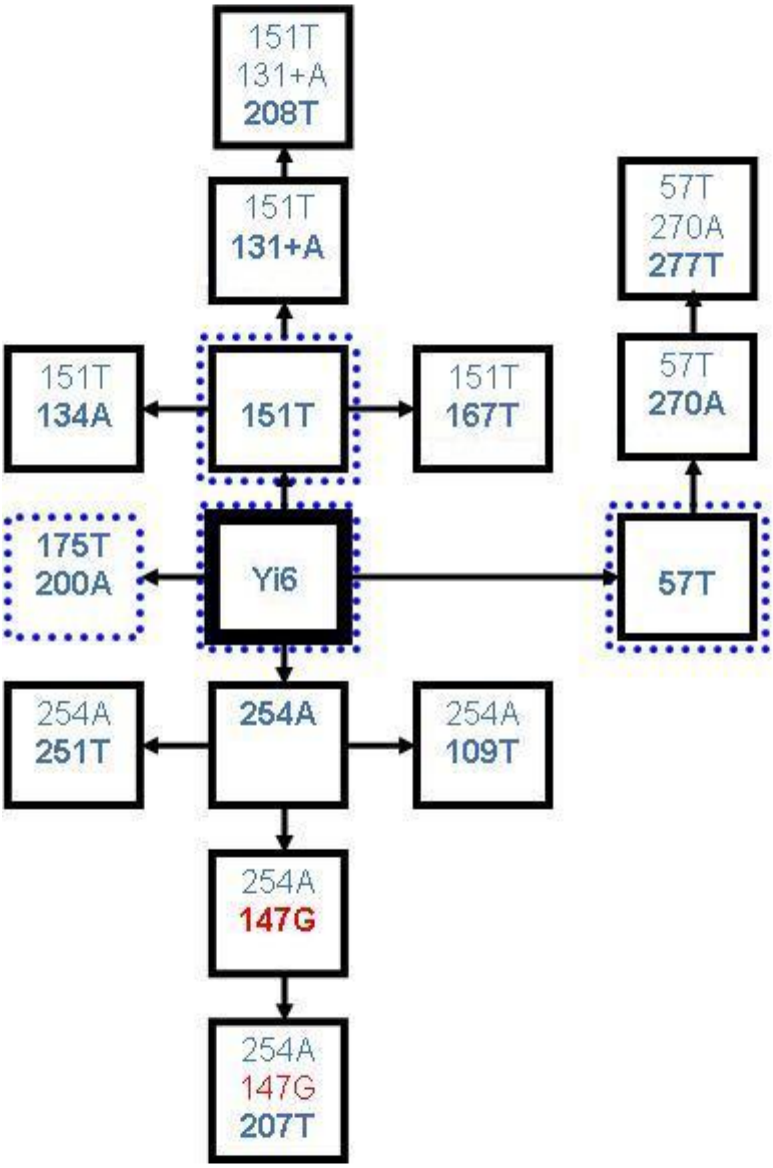


Figure 2.10: Inferred relationships between AluYi6 derivative subfamilies. Diagnostic mutations for each putative subfamily are shown in each box. The copy numbers of each of these putative subfamilies are listed in the Table 2.4. Blue dotted lines indicate the presence of a subfamily in chimpanzees. The mutation shown in red is a back mutation to the ancestral nucleotide at an AluYi6 diagnostic site.

Two of the elements that were found to be shared between humans and chimps in the original study were also found in the gorilla (Salem et al. 2003a), suggesting the subfamily arose before the divergence of gorillas from chimpanzees and humans, approximately seven million years ago. Yi6 does appear to have undergone some proliferation along the gorilla lineage, with at least one gorilla-specific element present in this species. AluYi6 appears to be absent from the orangutan, with no evidence of the subfamily in the orangutan pre-ensembl shotgun assembly. This suggests the subfamily is less than 10 million years old (Ackerman et al. 2002;Glazko and Nei 2003) .

Putative subfamily	Species	Mutations from Yi6 consensus	Copy number	Polymorphic?
Yi6	H, C, G	-	237 (H), 91 (C)	Yes
Yi6.1	H, C	57T	36 (H), 7 (C)	Yes
Yi6.1a	H	57T, 270A	10	Yes
Yi6.1b	H	57T, 270A, 277T	4	Yes
Yi6.2	H, C	151T	77 (H), 17 (C)	Unknown
Yi6.2a	H	151T, 134A	8	Yes
Yi6.2b	H	151T, 167T	5	Unknown
Yi6.2c	H	151T, 131+A	53	Yes
Yi6.3	H	151T, 131+A, 208T	22	Yes
Yi6.4	H	254A	35	Unknown
Yi6.4a	H	254A, 251T	5	Unknown
Yi6.4b	H	254A, 109T	3	Unknown
Yi6.4c	H	254A, 147G	20	Unknown
Yi6.4d	H	254A, 147G, 207T	18	Yes
Yi6.5	C	175T, 200A	31	Yes

Table 2.4: Putative AluYi6 derivative subfamilies. Yi6.1, Yi6.2 and Yi6.3 have been previously reported (Salem et al. 2003a). Where a shared mutation is found with additional shared mutations, the copy number of elements with the single mutation includes copies with further shared mutations. H = human, C = chimp, G = gorilla.

Gene conversion has also operated in the AluYi6 subfamily. There is evidence for a complete gene conversion event, as there is an AluYi6 element present in the chimpanzee (DQ59), and an older Alu belonging to the AluY subfamily present at the orthologous locus in humans.

2.4 Discussion

It appears that a young Alu subfamily can be defined in two ways, either by its present-day sequence, or by its ancestry (that is, its evolutionary history at its genomic locus since the moment of integration). A subfamily may be described as a collection of elements with the specified base at defined diagnostic positions. In this case, any Alu element with those diagnostic bases would be defined as a member of that subfamily. Alternatively, an Alu element can be defined as belonging to a subfamily if it is reasonable to assume that at the moment of integration, the sequence corresponded to that of the subfamily source gene, and may have since undergone gene conversion or back mutation such that it might show diagnostic changes. For example, it is feasible that many, if not all, of the elements presented here with, for example, only a single diagnostic change, integrated into the genome as elements with all correct diagnostic bases. Their inclusion in the subfamily acknowledges aspects of the evolutionary history of that particular subfamily that would otherwise be ignored. Therefore, for evolutionary analyses, the inclusion of an element within a subfamily based on its inferred state at the moment of integration seems more appropriate than inclusion based solely on its present-day sequence.

Groups of elements have been identified for several subfamilies, most notably AluYg6 and AluYi6, which do not appear to have been propagated by a source gene possessing the subfamily consensus sequence. This confirms the existence of “secondary” source genes within what has previously been considered a single subfamily. However, the idea of secondary source genes is poorly defined, as unless such source genes were identical to the original subfamily consensus, they would propagate diagnostic mutations themselves.

This would generate small collections of elements which can themselves be considered new subfamilies, as shown here, for example in the case of AluYg6a2 and AluYg5b3. It is not clear at what point a source gene with a mutation from the consensus of the subfamily from which it has arisen should cease to be regarded as a secondary source gene of its ancestral subfamily, and be considered a primary source gene and consensus sequence for a new derivative subfamily. Where a source gene can be seen to be producing daughter elements with unique mutations relative to the ancestral source gene, it is reasonable that these daughter elements be considered a derivative subfamily. Such subfamilies would still be considered as members of the ancestral subfamily for the purposes of evolutionary analyses, as in chapter 3. However, inclusion of these sequences in studies where the mutational variation from the subfamily consensus seen among the elements is used to make inferences about their evolution (for example, in estimating the age of the elements), would artificially inflate the total number of mutations seen, as some of these changes have been propagated by retrotransposition rather than mutation. This is due to the fact that at the time of insertion, elements derived from secondary source elements will already appear to have acquired mutations, and these elements will therefore appear to be older.

2.5 Conclusions

It is clear that there is considerable variation in the number of source genes present in each of the young Alu subfamilies. Evidence from patterns of shared mutations and polymorphism data suggest that multiple source genes are actively retrotransposing in the AluYh7, AluYg6 and AluYi6 subfamilies, the latter of which may contain up to 14 source elements. There are at least three active source genes within the AluYg6 subfamily, two of which have given rise to the new small subfamilies AluYg6a2 and AluYg5b3. There is not sufficient evidence to suggest the presence of secondary source genes contributing to the proliferation of AluYh3a1. The small AluYh3a3 subfamily appears to have followed the master gene model of proliferation in both humans and chimpanzees, with its substructure easily explained without the need to infer the activity of secondary source elements. Gene conversion appears to have operated in the AluYh3a1, AluYg6, AluYh3a3 and AluYi6 subfamilies, with partial gene conversion introducing ancestral mutations at diagnostic sites, and both forward and backward complete gene conversion replacing Alu elements with those belonging to other subfamilies. In the case of AluYh3a3, such an event has resulted in inactivation of the putative master gene in humans. The two small subfamilies descended from AluYg6 illustrate the ambiguity regarding Alu subfamily definition. Having access to the sequence data for a complete young Alu subfamily will be useful for exploring new computational methods for investigating the evolution of young Alu elements.

Chapter 3 - Simulating the evolution of young Alu subfamilies

3.1 Introduction

Given that Alu elements in humans and chimpanzees are rarely deleted, collections of Alu subfamilies, i.e. Alu elements that have derived from a particular source gene possessing specific diagnostic mutations, extracted from the complete genome sequences of these species can be assumed to represent entire subfamilies. Therefore, it should be possible to use the information stored within the copy number of each subfamily, and the shared and unique variation among those elements, to make inferences about the evolutionary history of each subfamily. Several *in silico* methods have been described which attempt to achieve this goal, such as the NETWORK approach (Cordaux et al. 2004), and the method of Hedges et al. (2005), which examines the level of insertion polymorphism and nucleotide diversity to make inferences about the evolutionary history of young Alu subfamilies. The latter model does, however, assume a master gene model of proliferation.

Here, a new *in silico* method for investigating the processes involved in the evolution of young Alu subfamilies is presented. The method is used to make inferences about the amplification histories of young Alu subfamilies AluYg6, AluYh7, AluYh3a1, and AluYi6, for example by determining likely values for parameters such as the number of source elements, i.e. the number of active Alu elements that have given rise to daughter elements, that have contributed to the proliferation of various subfamilies. A large parameter space is investigated to account for the unknown factors in young Alu subfamily proliferation, to minimise the number of assumptions required.

In this chapter, a new method for simulating the evolution of young Alu subfamilies is described, and used to investigate the number of source elements, and the retrotransposition rate, that may have been involved in the formation of the AluYg6 and AluYh7 subfamilies in humans, and the AluYi6 and AluYh3a1 subfamilies in both humans and chimpanzees. Firstly, the C++ program that was written to simulate young Alu subfamily evolution is described, including the file structure, input parameters, functions, variables, interface and assumptions. In the results and discussion, the effect of gene conversion in the program is discussed, followed by the results pertaining to the estimates of number of source elements for each subfamily generated by the program. Number of sources is estimated for the AluYh7 subfamily, followed by AluYg6, AluYh3a1 in humans and chimpanzees, and finally AluYi6 in humans and chimpanzees. The estimates for the rate of retrotransposition are then presented by subfamily, in the same order.

3.2 Methods

A C++ program (Appendix 2) was written to simulate the process of proliferation of young Alu subfamilies, starting with a single founder element and generating a complete subfamily in each case. Each run of the program generates a new subfamily, containing a predefined number of elements, according to certain parameters. The program then outputs features of the simulated subfamily, such as the number of source elements it contains. The subfamilies generated by each run consist of elements related to each other in different ways each time, due to the effects of random numbers generated as the program runs. The program compares the statistics, described below, of the subfamily generated in each run with those describing the real subfamily, which allows the proportion of runs for each parameter combination resembling the real data to be calculated. The average of each of the statistics describing each subfamily is also calculated. Therefore, the parameters generating those subfamilies, which, on average, most closely resemble the real subfamily, can be identified, as well as the total range of parameters which are able to generate subfamilies resembling the real data. This will therefore reveal the values for parameters, for example retrotransposition rate and number of source elements, which are most likely to, or could potentially, have operated in the evolution of a particular subfamily in reality.

The program consists of four C++ files, including two header files, Subfamily.h and Element.h. Subfamily.h defines a class called Subfamily, containing all the required functions and variables for the Subfamily.cpp file, such as the retrotransposition function, which relate to the construction of an entire subfamily. Element.h defines a class called Element, containing all the required functions and variables for the Element.cpp file, such as the mutation function, which relate to individual elements within the subfamily. Subfamily.cpp contains the main function, where all events required to produce an entire subfamily from a single element occur.

Each element in the subfamily is an object with the following attributes: designation, sequence, parent, active, source, and number of mutations. These attributes function as follows:

Designation: increments by one when a new element is added, so the designation of each new element is one greater than the previous element generated.

Sequence: the sequence of element 1 corresponds to the subfamily consensus, as this is assumed to be the ancestral sequence of the subfamily founder element. Each new element inherits the sequence of its parent at the time of its creation.

Parent: the designation of the element which provided the template for the new element.

Active: a Boolean variable describing whether or not the element is retrotranspositionally competent, determined when the element is created. An element has a probability, pA, of being active (discussed below), and activity is assigned randomly to elements according to this probability.

Source: a Boolean variable describing whether or not the element has generated any new elements by retrotransposition. A source element must be “true” for the active attribute.

Number of mutations: a count of how many point mutations have occurred in the element relative to the consensus. The number of mutations is inherited from the parent and continues to be incremented. Element 1 is constrained and cannot mutate. There would be no reason for founder elements to be constrained in reality, but it is assumed that the subfamily consensus and the sequence of the founder element are the same. If mutations occurred in the founder element at an early stage, the subfamily consensus would be different.

A function can be called to view these attributes for the elements belonging to an entire subfamily as it is generated, and therefore precise relationships between elements can be determined.

For the program to run, the user must input values for the following parameters:

pA, pT and pG: the probability that a new element created is active (pA), the probability of retrotransposition per active element (pT) and gene conversion (pG) per year are unknown, so a range of values can be input to be tested. The size of the increment can also be input. For example, the user can instruct the program to test all values of pA between 0 (the master gene model) and 1 (the transposon model) in increments of 0.001. For this study, parameter values were tested for ranges on a logarithmic scale. For example, for pA, all values between 0 and 0.001 were tested with an increment of 0.0001, and then all values between 0.001 and 0.01 with an increment of 0.001. Parameter space between 0 and 1 was investigated for both pA and pT. pT increases as the number of active elements increases, with pT being equal to starting pT multiplied by the number of active elements. Therefore, the parameter space in which successful runs were obtained tends to range from small pT with large pA, to large pT with small pA. pG increases every time a new element is created, with pG equal to starting pG multiplied by the total number of elements.

pM: the probability of mutation per year, pM, can be input by the user. For this work, the probability of mutation used is the mutation rate of primate intervening DNA sequences adjusted for the proportion of CpG dinucleotides in each subfamily, which are assumed to have a rate of mutation six times that of non-CpG residues (Hedges et al. 2004). pM increases in the same way as pG, with pM equal to starting pM multiplied by the total number of elements. With the mutation rate for CpG dinucleotides six times higher than other residues, and a limited number of CpG sites, under certain values of pA and pT, all CpG residues become mutated to either CpA or TpG, such that CpG transition mutations can no longer occur. Under these conditions, the program does not complete the construction of the subfamily, and the parameter combination is deemed unsuccessful. This can be reported in the program output. Using parameter combinations for which this occurs for a proportion of the runs performed will result in fewer runs being used to generate the results for that combination. However, as such parameter combinations are highly unlikely to generate any results resembling real subfamilies, this is unlikely to be a problem. If such a parameter combination were to generate at least one successful run, however, rejecting runs in which all CpG sites had been mutated may introduce some bias into the estimate of factors such as the average number of source elements for that parameter combination.

The simulations are conducted in years, rather than generations, as they involve a single genome, rather than a population, within which a family of transposable elements is created by stochastic forces (retrotransposition, mutation and gene conversion), which occur with a constant probability. As only a single genome is simulated, the model does not include any of the factors involved in population genetics, such as genetic drift and changing frequencies of particular variants, which would require the program to be conducted in terms of generations rather than years. It would be possible, assuming a known generation time (e.g. of 25 years for humans), to convert the rates of retrotransposition, gene conversion and mutation, into units per generation rather than per year, which would have no effect on the outcome. It should be borne in mind that the generation time for chimpanzees is shorter than for humans, however, this difference in generation time is assumed to have no effect on the rates of the stochastic forces in the program.

Total number of elements: the number of elements in the complete subfamily.

Number of runs: for each run, the program produces a complete subfamily for each possible combination of parameters input by the user. Each subfamily produced in a run, although generated using the same parameters, will differ by chance due to the effect of random numbers generated in the program. For this work, 100 runs were conducted with each parameter combination. This number of runs was chosen as a compromise between the time constraint and the need to conduct enough runs to observe meaningful results. Inititial testing of the program to determine an appropriate number of runs which fulfilled these criteria involved performing 10, 50, 100, 500 and 1000 runs for a small group of parameter sets, and observing the distribution of results obtained for the number of source elements. It was anticipated that an ideal number of runs would result in a normally-distributed number of source elements for each parameter combination. It was found that 1000 runs were required to get an almost perfect bell curve, but the intitial testing revealed there was not sufficient time to perform 1000 runs for each of the parameter combinations to be tested. Some deviations from the bell curve were observed when 100 or 500 runs were performed, and a normally distributed set of results was not obtained with fewer than 100 runs. It was decided that 100 runs would provide an accurate enough result in the available time. By conducting 100 runs, it was possible to investigate a large parameter space in relatively small increments.

Subfamily name: the user selects the name of a subfamily, which retrieves the corresponding subfamily consensus, which is set as the value of the sequence attribute of element 1. The subfamily name also determines which sites will be considered “diagnostic” for use in the gene conversion function, and the length of the consensus sequence, which varies slightly between young Alu subfamilies.

The program is able to use a combination of four statistics (time, theta, pi and shared mutations) to decide whether or not a simulated subfamily closely resembles the real subfamily. Where a generated subfamily has values for a combination of these statistics within a defined range it is considered successful. Only details of successful runs are output by the program, to limit the number of results produced to those which closely resemble the real data, within the range defined by the user.

The user inputs a range of time which is considered realistic for the subfamily being simulated. For example, the AluYi6 subfamily is known to be present in humans, chimpanzees and gorillas (Salem et al. 2003a) but not in orangutans. An appropriate range would therefore be 7 – 10 million years, based on the known divergence times of these species. All subfamilies tested were considered to be currently active due to the presence of elements polymorphic for presence or absence, therefore it is not necessary for the program to account for proliferation having occurred in the past and then ended. The time therefore corresponds both to the time of onset of proliferation (mya), and the total time of proliferation (myr).

Theta, which represents the amount of variation expected at each site if evolution is neutral, was calculated for the real data using a simple C++ algorithm. Pi, a measure of nucleotide diversity corresponding to the average number of nucleotide differences between two random sequences in the subfamily, was calculated for the real data using DnaSP (Rozas et al. 2003). The user is able to input a range of values within which theta and pi must fall in order for a run to be considered successful. For this study, values within 10% on either side of the real values for theta and pi for the subfamily were chosen.

Shared mutations are the number of non-unique mutations present in the subfamily. In the program, this is the sum of the number of parallel mutations, ancestral mutations at diagnostic sites introduced by partial gene conversion, and “shared source mutations”. Shared source mutations are mutations shared by descent, i.e. the total number of mutations that exist due to retrotransposition. Shared mutations will be large where there has been a considerable amount of parallel mutation, and where there are many secondary source elements. Again, the shared mutations statistic was required to be within ±10% of the real value, rounded down at the lower end, and up at the upper end, to the nearest whole number.

The functions in the program describe the processes which affect the evolution of the subfamily, including retrotransposition, gene conversion and mutation. The functions work as follows:

Event Decision: This function calls the functions retrotransposition, gene conversion and mutation, with differing probabilities based on the values entered by the user for the variables pT, pG and pM. Each event occurs a certain amount of evolutionary time after the last one, which is then added to the total time. This amount of time decreases as the number of elements increases. The time to the next event is determined by a random exponentially distributed number with a mean of 1, multiplied by 1/(pM+pG+pT). Time in the program is the time since the onset of proliferation for a particular subfamily, not the time since the origin of the subfamily, as the founder element may be dormant for some time after its generation (Han et al. 2005). Therefore, time starts after the first event, which is always a retrotransposition event. The program runs, and events continue to happen, until the appropriate number of elements for each subfamily has been generated.

Activation: when a new element is created, the value of its “active” attribute must be set to true (if it is to be retrotranspositionally competent) or false (if it is incapable of retrotransposing). The probability of an element being active is determined by the user input variable pA (probability of activation). A random number between 0 and 1 is generated, and if this value is less than or equal to the value of pA, active is set to true, otherwise it is set to false.

Retrotransposition: an element for which the value of the attribute "active" is true is selected at random. A copy is made of this element, creating a new element object with a designation one greater than the previous object. The new element inherits its sequence and number of mutations from its parent, and its active status is determined by calling the activation function. A variable to count the total number of elements is incremented by one. The probability of retrotransposition occurring is determined by the value of user input variable pT relative to the values for pM and pG. The probability that the next event is a retrotransposition, rather than a mutation or gene conversion, is pT/(pT+pM+pG). The probability of the next event being a mutation or gene conversion is calculated in the same way.

Mutation: if a mutation is to occur, first an element is chosen to be mutated, then, a type of mutation is selected. This is according to the different probabilities of different types of mutation occurring, i.e. CpG transitions, non-CpG transitions, transversions at CpG sites and transversions at non-CpG sites. CpG transitions occur at six times the rate of other mutations (Xing et al. 2004). Then, a site is chosen to be mutated. The two possible nucleotides generated by a transversion mutation have an equal probability of being chosen.

Gene conversion: this function calls three other functions with differing probabilities. These functions represent the nature of gene conversion, in that it can be complete, where an Alu element from one subfamily is completely converted to another, or partial, where only a short stretch of sequence within the element is converted. In the program, 40% of gene conversion events in the program are complete, and 60% are partial.

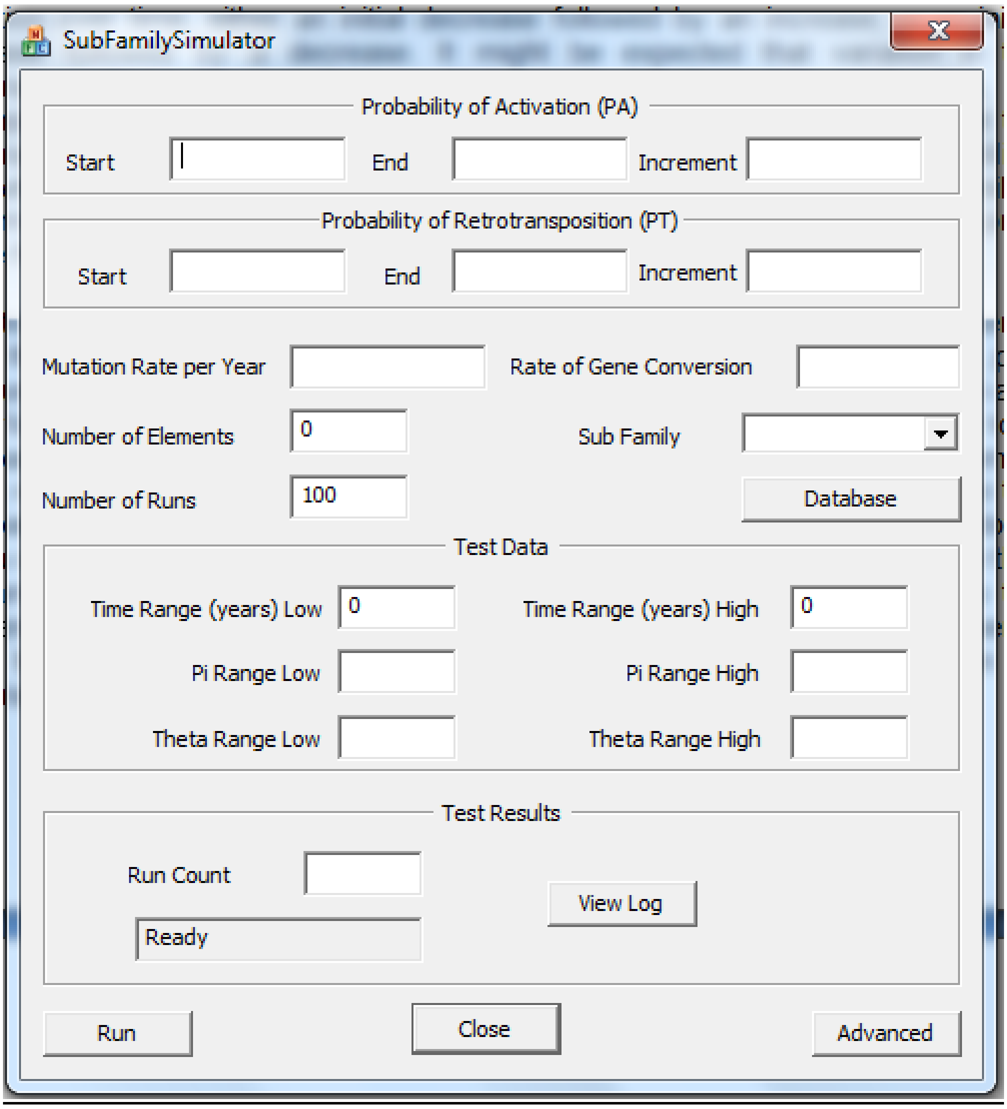
Complete Forward: Where complete gene conversion occurs, half of the time this will be a forward gene conversion event, which generates a new element. This reflects the assumption that in reality, where complete gene conversion occurs, the two elements involved would be equally likely to provide the template for the conversion. This function behaves like the retrotransposition function, by making a copy of an existing element, however, in this case, the parent can either be active or inactive. The parent, sequence, designation and number of mutations are set as for the retrotransposition function.

Complete Backward: half of complete gene conversion events are “backward”, converting an element of the subfamily to one of an older subfamily. For the purpose of the program, this simply requires that an element is selected at random and deleted.

Partial: There are two templates available to the program for partial gene conversion – AluSg and AluY. AluY is chosen as a template more frequently than AluSg, at a ratio of 2:1, due to the higher similarity between young Alu elements and AluY. Partial gene conversion tracts in Alu elements are generally assumed to be around 50-100bp in length (Batzer and Deininger 2002;Roy-Engel et al. 2002). This function chooses a position as the start of the gene conversion tract, and then a tract length between 50 and 100. The end position is therefore the tract start position added to the tract length, unless this would exceed the end of the element, in which case the tract stops at the end of the element. The corresponding region is then copied from one of the templates and pasted into the sequence of the selected element, replacing the existing sequence in the region covered by the gene conversion tract.

To determine whether allowing for the rate of retrotransposition to vary over time improved the ability of certain parameter sets to generate subfamilies resembling the real data, the event decision function was modified to allow the user to input either a linear or quadratic fluctuation in retrotransposition rate over time. Instead of the retrotransposition rate consistently corresponding to pT, the user is able to input a value for a gradient (m) and y-intercept (c) for a linear increase or decrease in rate over time. Alternatively, values for a, b and c can be input to generate a quadratic variation over time, either an initial decrease followed by an increase, or an initial increase followed by a decrease. It might be expected that variation in the retrotransposition rate over time would be applicable to certain subfamilies, for example, those that are present in both humans and chimps. It is known that the retrotransposition rate has increased in humans relative to chimpanzees (Mills et al. 2006), and therefore for these families in humans, a linear increase in rate might be appropriate, to reflect a slow rate prior to the human-chimp divergence, then increasing along the human lineage.

An interface was designed (Bob Scarle, personal communication) to allow the user to input the range of values for pA and pT to be tested, along with values for pM, pG, the number of elements in the subfamily, and the range of values for theta, pi and time for which a run is to be considered "successful" (Figure 3.1). The interface is linked to a database of Alu subfamilies, which contains the consensus sequence for each family, along with the diagnostic positions for that family. New families can be added to the database. The user can select a subfamily from the database using a drop-down menu on the interface, which will then incorporate the necessary information for that subfamily into the program, for example, using the consensus sequence as the sequence for the founder element of the subfamily. In the advanced tab, the user is able to input values to assign either a linear or quadratic fluctuation in retrotransposition rate over time.



Number of Elements: 0

Number of Runs: 100

Time Range (years) Low: 0

Time Range (years) High: 0

Ready

Figure 3.1: Interface of the SubfamilySimulator program, allowing the user to input all required parameters.

The program makes several assumptions about Alu evolution. In the program, elements can be obscured by complete backward gene conversion, but it is not possible for elements to be deleted. As deletion of Alu elements is rare (Belle et al. 2005;van de Lagemaat et al. 2005), particularly for young Alu subfamilies, this possibility if not incorporated into the program. Likewise, insertions and deletions within Alu elements are also ignored, as they are relatively rare. For example, only three mutations of this nature are observed in the AluYh7 subfamily, two of which are expansions of the central A-rich region. In the AluYi6 subfamily in humans, which contains 237 elements, only 10 small insertions and deletions (1-3bp), other than expansions of the central A-rich tract, are observed. It is assumed that all active elements within a subfamily retrotranspose at the same rate. Therefore, each active element is equally likely to be selected by the retrotransposition function. Active elements are also assumed to remain active throughout the generation of the subfamily. The rates of mutation and gene conversion per element are constant through time for each subfamily, as is the rate of retrotransposition unless a fluctuation is applied by the user. Initial testing revealed that modification of the retrotransposition rate over time did not improve the results obtained for any of the subfamilies tested, therefore, in this study, pT is constant over time. There is assumed to be no mutation rate variation along the length of the sequence. Therefore, each CpG site has the same probability of mutating, as does each non-CpG site. Finally, the possibility of 5' truncation upon integration is not included in the program. Although 5' truncation events are relatively common for Alu elements, these events are assumed not to have a substantial impact on the statistics being assessed.

The values for each of the parameters used in the program are shown in Table 3.1. Two parameters are consistent for all subfamilies: gene conversion is assumed to be negligible and therefore pG=0 (see results); and the mutation rate is also fixed, with pM=7.95 x 10^{-7}. This is based on the mutation rate for primate intervening DNA sequences adjusted for the presence of CpG dinucleotides mutating six times faster than other sites (Xing et al. 2004). The values for theta and pi for the real subfamilies are given.

Subfamily	Number of elements	Onset of proliferation (mya)	Theta	Pi	Shared mutations
AluYh7	20	0 - 4	0.01431	0.00593	2
AluYg6	380	3.5 – 5.5	0.111263	0.02173	787
AluYi6 (human)	237	7 – 10	0.104138	0.03271	641
AluYi6 (chimp)	91	7 – 10	0.0855029	0.03116	177
AluYh3a1 (human)	98	10 – 16	0.0822143	0.03955	228
AluYh3a1 (chimp)	73	10 – 16	0.0964648	0.05446	292

Table 3.1: Statistics for the subfamilies investigated using the program.

3.3 Results and Discussion

The program was used to determine values for the retrotransposition rate per element, and the number of source elements, which are likely to have generated the subfamilies tested, in addition to the total range of possible values. Where the number of source elements was found to be 1, a master gene model can be inferred. Otherwise, an intermediate model, where multiple source elements are present, can be inferred.

3.3.1 Gene Conversion

Values for pG, the probability of gene conversion per year, with the full range of possible values for pA and pT, between 0 and 1, were tested to determine whether inclusion of gene conversion improved the outcome of the simulations in terms of generating subfamilies which closely resembled the real data. For each combination of parameters, inputting a probability of gene conversion did not consistently generate results that were, on average, a better fit for the real data, relative to setting pG to 0, and the percentage of successful runs per parameter combination in many cases decreased, and in others did not significantly increase. This might suggest that gene conversion has not been very influential in the evolution of the Alu subfamilies tested. Therefore, after initial testing, pG was set to 0; thus the results generated are based on the assumption that there has been no gene conversion. Testing the range of values for pG revealed that inclusion of gene conversion did not significantly alter the estimates of number of source elements for each parameter combination (Table 3.2). It may be expected that the number of sources required would be smaller due to the effect of forward gene conversion events generating new elements, but as backwards gene conversion events occur at the same rate, eliminating elements, this does not occur. It may also be expected that partial gene conversion events may increase the number of shared mutations between elements, which is also likely to decrease the number of sources required to account for the number of shared mutations observed. This effect was not consistently observed, and therefore the results presented below, which assume gene conversion has not occurred, can be assumed to be reliable in terms of the estimation of number of source elements and retrotransposition rates. However, initial testing indicates that the program cannot successfully be used to confidently determine the most likely rate at which gene conversion has operated on these subfamilies, as a wide range of values for pG generate successful runs.

pG	Percent success				Average sources				Average shared mutations			
0	8	9	16	7	2.53	2.56	2.43	2.58	15.84	7.38	6.24	4.41
$1x10^{-10}$	8	10	8	12	2.74	2.64	2.61	2.57	14.06	8.06	6.23	5.97
$1x10^{-9}$	8	9	7	12	2.66	2.65	2.54	2.55	13.57	8.36	5.98	6.35
$1x10^{-8}$	7	10	7	13	2.49	2.61	2.65	2.64	14.40	8.35	6.93	5.27
$1x10^{-7}$	13	12	13	9	2.59	2.53	2.49	2.56	18.68	10.49	6.10	4.29

Table 3.2: Variation in the percent success, average number of source elements, and the average number of shared mutations for different values of pG, for the parameter combination pA = 0.1, pT = $2x10^{-6}$ - $5x10^{-6}$ for the AluYh7 subfamily. These four values are given in the four columns in each of the "Percent success", "Average sources" and "Average shared mutations" columns.

The lack of ability of the inclusion of gene conversion to produce results more closely resembling the real data may be due to the fact that the templates used in the partial gene conversion function are consensus sequences, and therefore do not contain mutations. In reality, most older Alu elements, particularly AluS elements, contain many mutations from their consensus, and these would be introduced by partial gene conversion. In the program, undergoing partial gene conversion will introduce ancestral nucleotides at diagnostic positions, and if AluSg is used as a template, additional mutations corresponding to the diagnostic mutations of the AluSg subfamily. However, it will not introduce additional mutations, and may actually eliminate variation which has been generated in an element throughout the course of the program due to mutation. This would be particularly true if the partial gene conversion function were called relatively late in the formation of the subfamily, as there would have been sufficient time for older elements to become quite diverged from the consensus. This may account for the inclusion of gene conversion producing subfamilies which resemble the real data to a lesser extent than when gene conversion is excluded.

For AluYg6, the number of shared mutations obtained by the program is generally much lower than the real value. This might suggest that gene conversion has had a substantial impact on AluYg6 evolution, introducing large numbers of mutations. This is also supported by the excess of ancestral nucleotides at diagnostic positions in this subfamily (Styles and Brookfield 2007).

3.3.2 Number of Source Elements

A summary of the number of source elements estimated from the sequence data, polymorphism data and the program for each subfamily tested are shown in table 3.6 (page 119). The number of sources estimated from the sequence data is equivalent to the number of sets of mutations shared by a considerable number of elements in the subfamily, which are assumed to be shared by descent. The polymorphism estimate corresponds to the number of sets of shared mutations for which elements possessing those mutations are polymorphic for presence or absence. The presence of shared mutations in a polymorphic element indicates the mutations are more likely to be shared by descent than due to parallel mutation, even in cases where the number of elements sharing the mutations is very small. The number of source elements estimated from polymorphism data may, in some cases, be lower than the actual number of source elements that have contributed to the proliferation of the subfamily, in cases where source elements are no longer active. It may also be an underestimate due to the limitations of using the trace archives to estimate levels of polymorphism.

Each combination of parameters generated 100 subfamilies, which differed due to the effects of random numbers generated in the program. The mean was then taken of the values of all statistics in the 100 runs to give an average value for theta, pi, proliferation time, shared mutations and number of source elements for a given combination of input parameters. The optimal set of parameters is defined as that which gives an average, across 100 runs, for theta, proliferation time and shared mutations, most closely resembling the real subfamily. Values of pi produced by the program were generally lower than the real values, therefore were not included in the assessment of optimal parameter combinations. A possible explanation for the small values for pi is that, in reality, partial gene conversion events may introduce many mutations which would inflate the value of pi, which does not occur in the program. A single run of the program was considered successful if the value for theta for the population of elements produced in that run fell within the required range, which was within 10% on either side of the real value. The proliferation time was also required to fall within the range given in Table 3.1. Therefore, in addition to determining the parameter values and number of sources which generated subfamilies which, on average, most closely resembled the real data, the most successful parameter combination could be determined, that is, the combination which generated the greatest proportion of successful runs. This analysis was done under three conditions: inclusion of only theta and proliferation time as success tests (condition 1), inclusion of theta, time and shared mutations (condition 2), and finally inclusion of theta, time, and pi (condition 3).

The total possible range of values for the parameters pA and pT, in addition to the number of source elements, was determined. This corresponds to the range of parameters which were able to generate at least one successful run, as these can be assumed to be theoretically capable, although less likely, to generate a subfamily resembling that seen in reality. For each subfamily, the number of sources, and the retrotransposition rate, is reported in five ways. Firstly, the “optimal” number of sources, that is, the average number of sources in the parameter combination that generates subfamilies which, on average, most closely resemble the real subfamily in terms of theta, proliferation time and number of shared mutations. The “most successful” refers to the average number of sources in the parameter combination which generates the greatest proportion of successful runs. The “most frequent” corresponds to the number of sources most frequently observed in successful runs. The “average” is the average number of sources observed in successful runs only, and finally, the “total possible” number of sources corresponds to the total range of number of sources observed in runs which generate subfamilies resembling that seen in reality.

In the following sections, graphs are presented to illustrate the number of source elements estimated in all successful runs, regardless of the parameter values required to generate each run, i.e. the frequency of each number of sources in the “total possible” range. The “most frequent” number of sources can also be observed, as can the distribution of all possible number of sources across all successful runs. As the parameter combinations that reflect the real-life evolution of the young Alu subfamilies are unknown, although the parameter combinations which generate an average most closely resembling the real data, or those with the highest percentage success, may be deemed most likely to apply to the generation of the subfamilies in reality, this is uncertain. These graphs therefore provide an indication of how likely a particular number of sources is to have generated the real subfamily, without assuming a particular parameter combination. Consequently, however, this presentation of the results does not allow for interpretation of which parameter combinations generate which values, and whether or not, for example, particular numbers of sources are observed only in a clustered range of parameter combinations, or whether these are spread more widely across the parameter space investigated.

Subfamily	Number of sources			
	Most successful	Most frequent	Average	Total possible
AluYh7	1.01	1	1.94	1 - 13
AluYg6	1	1	4.82	1 – 45
AluYh3a1 (human)	29.9	1	12.68	1 – 52
AluYh3a1 (chimp)	1	1	1.22	1 – 20
AluYi6 (human)	11.47	1	22.74	1 - 124
AluYi6 (chimp)	1.03	1	1.38	1 - 27

Table 3.3: Number of source elements required to generate subfamilies resembling those observed in reality estimated by the program under condition 1.

Subfamily	Number of sources			
	Most successful	Most frequent	Average	Total possible
AluYh7	1.05	1	1.092	1 - 6
AluYg6	19.44	4	16.5	2 – 41
AluYh3a1 (human)	4.01	1	3.26	1 – 12
AluYh3a1 (chimp)	3.43	1	2.01	1 – 8
AluYi6 (human)	8.72	1	9.74	1 – 28
AluYi6 (chimp)	1.1	1	1.26	1 - 6

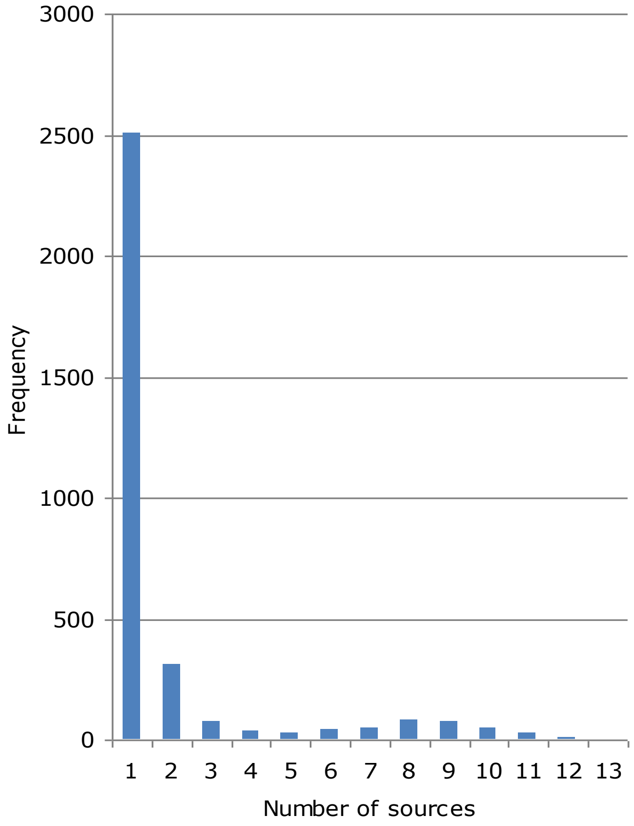
Table 3.4: Number of source elements required to generate subfamilies resembling those observed in reality estimated by the program under condition 2.

Subfamily	Number of sources			
	Most successful	Most frequent	Average	Total possible
AluYh7	-	-	-	-
AluYg6	19.44	19, 23, 26	26.06	3 - 105
AluYh3a1 (human)	17.84, 22.06	1	16.81	1 – 52
AluYh3a1 (chimp)	3.43, 9.44	1	4.45	1 - 13
AluYi6 (human)	47.06	13	34.30	1 - 120
AluYi6 (chimp)	5.8	1	2.35	1 - 17

Table 3.5: Number of source elements required to generate subfamilies resembling those observed in reality estimated by the program under condition 3. No successful runs for the AluYh7 family possessed a pi value within the required range.

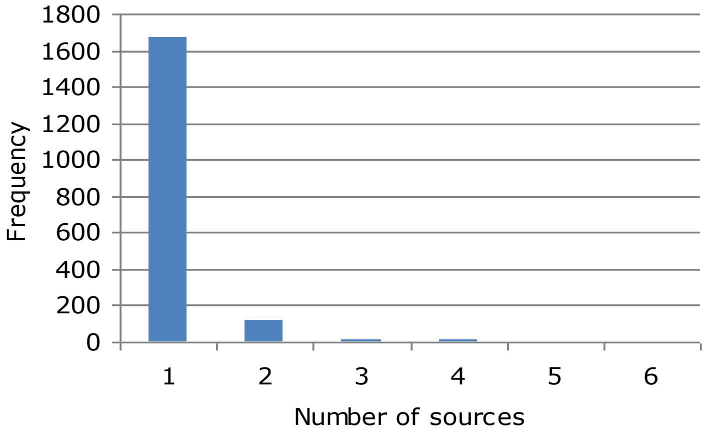
3.3.2.1 AluYh7

The most successful parameter combination, pA = $4x10^{-4}$ and pT = $6x10^{-6}$, as expected, has a small value for pA, such that few elements under this parameter combination will be active. This most successful combination generated 35 out of 100 runs resembling the real data under condition 1, with an average number of sources of 1.01 (Table 3.3), with 99 of the runs possessing one source element, and a single run possessing two. Under condition 2, the parameter combination which was the second most successful under condition 1, pA = $3x10^{-3}$, pT = $5x10^{-6}$, is most successful, with 26% of runs in the required range. This parameter combination, with its larger value of pA, increases the most successful number of sources to 1.05 (Table 3.4). The average number of sources across 100 runs for the parameter combination which generates subfamilies which, on average, most closely resemble AluYh7 is 1.54. The most frequent number of sources in successful runs is 1, found in 2510 out of 3283 (76.5%) successful runs under condition 1, and 1669 out of 1804 (92.5%) under condition 2, suggesting that the AluYh7 subfamily could have been generated under the master gene model. The average number of sources under condition 1 was 1.93, which falls to 1.092 under condition 2. Pi did not fall within the required range in any successful runs for AluYh7 (Table 3.5). The total possible number of sources able to generate the AluYh7 subfamily under condition 1 was between 1 and 13 (Figure 3.2), which fell to 1 – 6 under condition 2 (Figure 3.3).



Source	Frequency
1	2510
2	312
3	76
4	34
5	27
6	42
7	46
8	83
9	73
10	47
11	27
12	5
13	1

Figure 3.2: The frequency of each possible number of source elements in successful runs for the AluYh7 subfamily, under condition 1, summed across all runs.



Source	Frequency
1	1669
2	114
3	14
4	5
5	1
6	1

Figure 3.3: The frequency of each possible number of source elements in successful runs for the AluYh7 subfamily, under condition 2.

3.3.2.2 AluYg6

For AluYg6, there are three combinations of parameters which, on average, give equally good matches to the real subfamily. These give average numbers of sources across 100 runs of 3.19, 3.61 and 3.98. The average of these three values, 3.59, is taken as the optimal number of sources for AluYg6. This value is consistent with the sequence data, which suggest three or four source elements based on patterns of shared mutations (Styles and Brookfield 2007). The polymorphism data suggest there may be six source elements in this subfamily. As two of these potential source genes have produced very low numbers of elements, it is unsurprising that the optimal number of sources in the simulations is lower than that suggested by the polymorphism data (Table 3.6). The sixth source element suggested by polymorphism data is a case in which 23 elements share a single mutation in addition to the AluYg6 diagnostic mutations (Styles and Brookfield 2007), which can be referred to as

AluYg6c1. At least six (approx. 26%) of these AluYg6c1 elements are polymorphic.

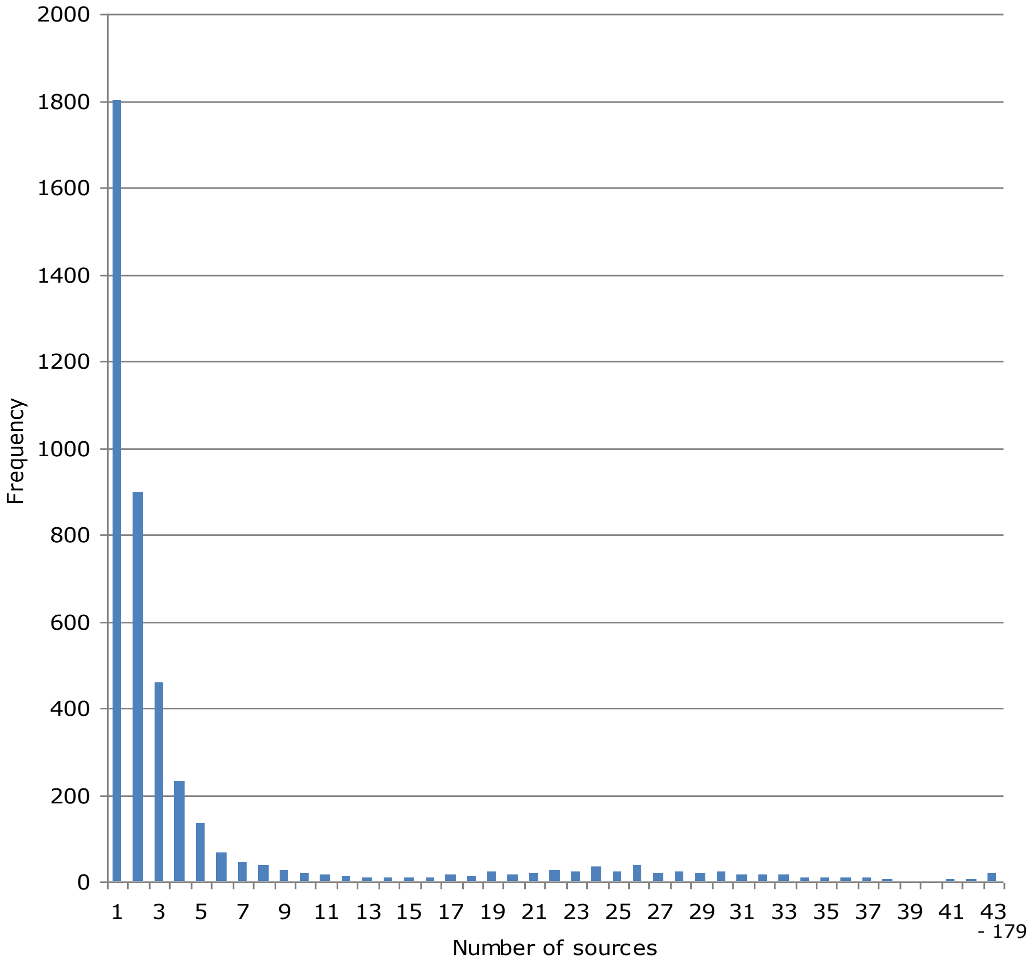


Figure 3.4: The frequency of each possible number of source elements in successful runs for the AluYg6 subfamily, under condition 1.

The most frequent number of sources observed in successful runs for AluYg6 under condition 1 is 1 (Figure 3.4). However, as many successful runs possess greater than one source element, AluYg6 may have been formed by the activity of multiple source elements, rather than under the master gene model. However, the most successful parameter combination, with 72% of runs falling within the required range, also had only one source element in each run, with a pA value of 0, and pT of 1x10^{-4} (Table 3.3). The high frequency of single source elements generating successful runs, despite polymorphism data suggesting the presence of multiple source elements, might suggest that a single source element has a much greater level of activity than the secondary source elements. The program makes the assumption that if a retrotransposition event occurs, each of the active elements present is equally likely to be copied. Under condition 2, generally, a much larger number of sources is required to generate the high number of shared mutations observed for AluYg6, in the absence of gene conversion. Under condition 2, the range of sources capable of generating the AluYg6 subfamily is greatly restricted compared to condition 1, and no longer allows the master gene model, as the number of shared mutations observed is too high (Table 3.4). The most frequent number of sources is 4 (Figure 3.5), in line with estimates from sequence data, found in 16 out of 166 (9.6%) successful runs. The most successful parameter combination, with 9 out of 100 runs within the required range, is pA=0.05, pT=1x10^{-5}, which generates an average of 19.44 sources. This is again the most successful combination under condition 3 (Table 3.5), with 7% success. 19, 23 and 26 sources are equally frequent among the successful runs under condition 3, each found in 9 out of 140 runs. The average of all successful runs under condition 3 is once again higher, at 26.06, and the total possible range once again increases to between 3 and 105 (Figure 3.6), excluding the master gene model as under condition 2.

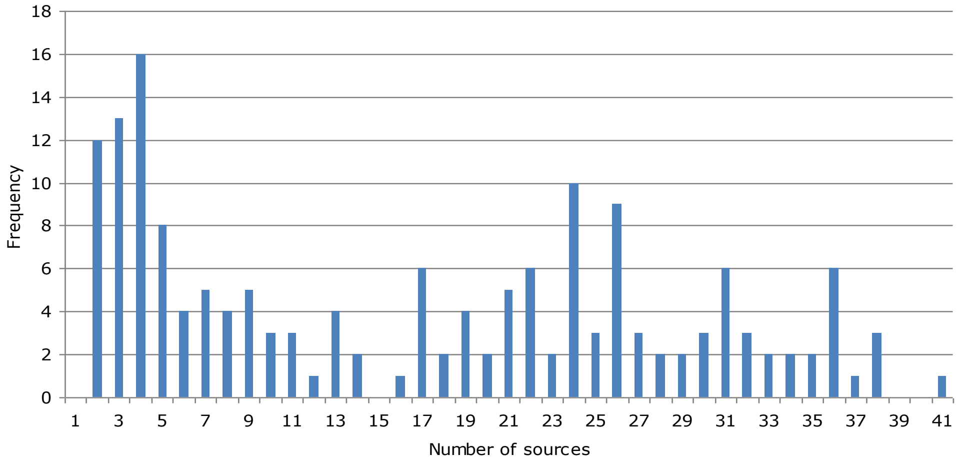


Figure 3.5: The frequency of each possible number of source elements in successful runs for the AluYg6 subfamily, under condition 2.

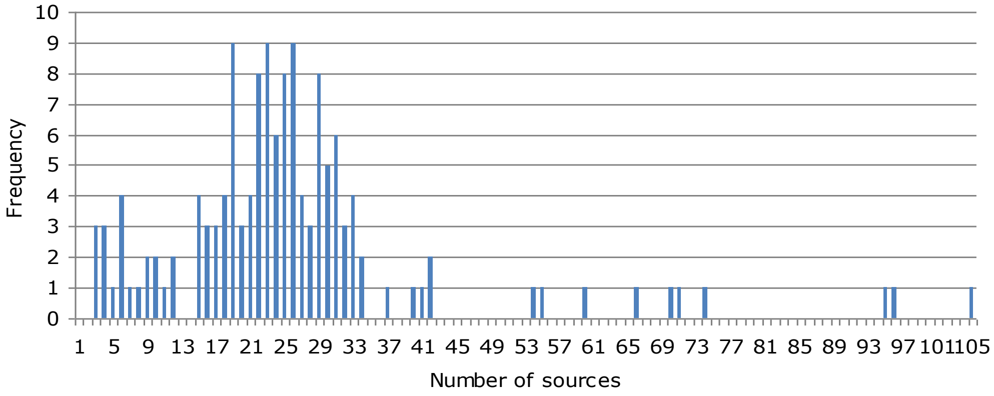


Figure 3.6: The frequency of each possible number of source elements in successful runs for the AluYg6 subfamily, under condition 3.

3.3.2.3 AluYh3a1

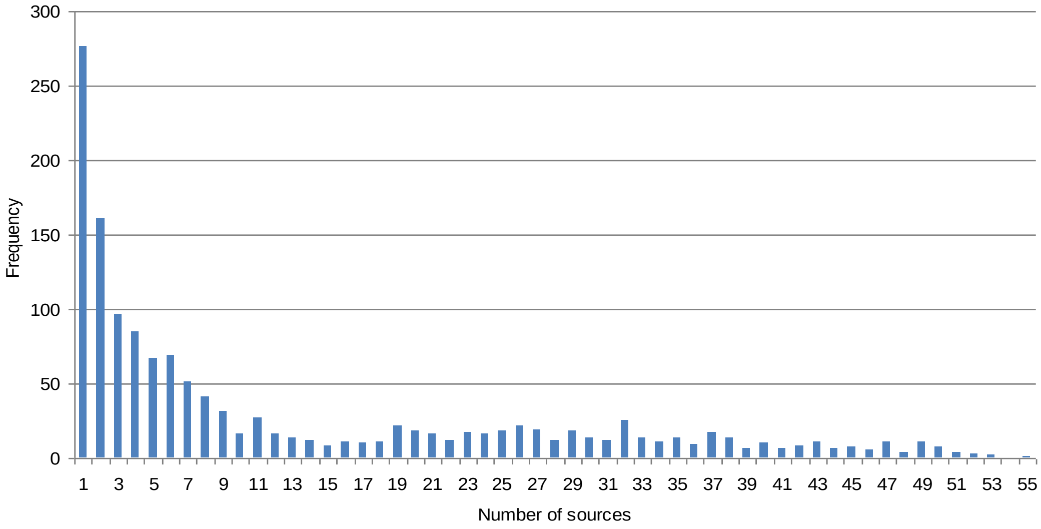


Figure 3.7: The frequency of each possible number of source elements in successful runs for the AluYh3a1 subfamily in humans, under condition 1.

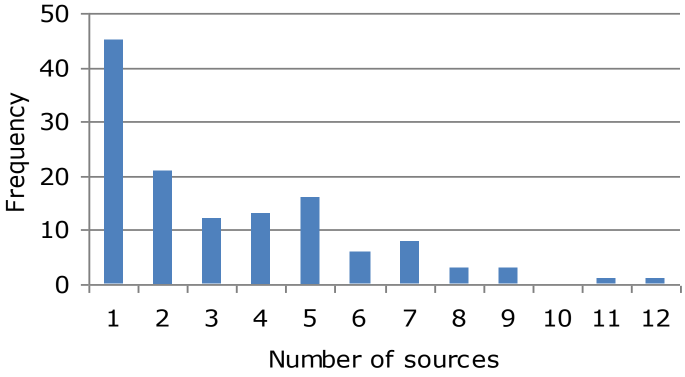


Figure 3.8: The frequency of each possible number of source elements in successful runs for the AluYh3a1 subfamily in humans, under condition 2.

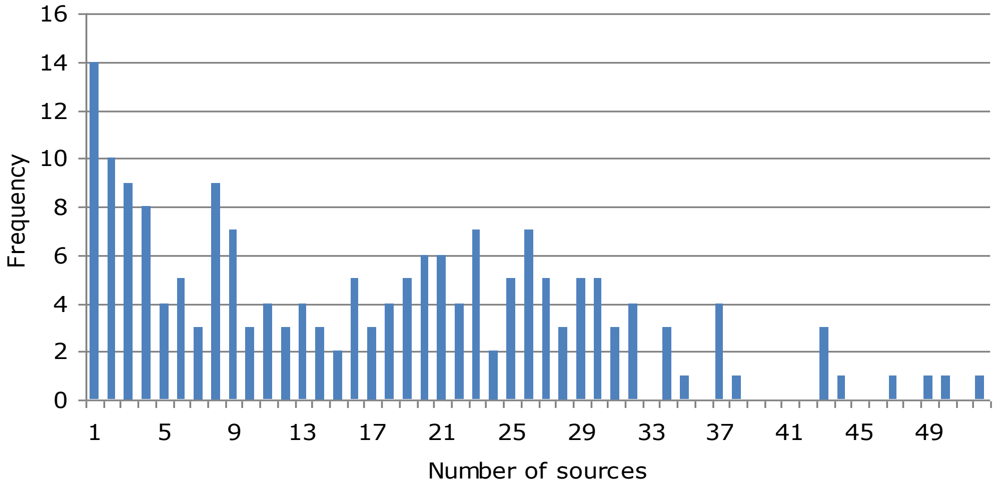
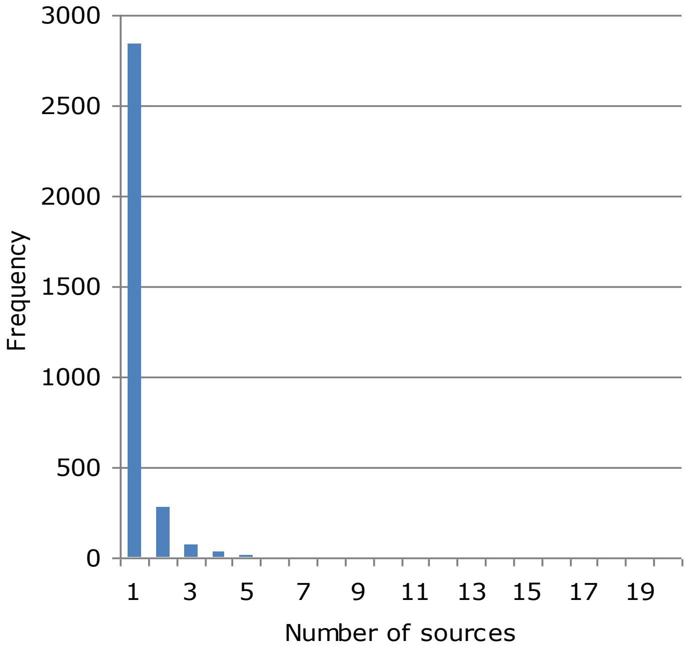


Figure 3.9: The frequency of each possible number of source elements in successful runs for the AluYh3a1 subfamily in humans, under condition 3.

There is no outstanding evidence from the sequence data for secondary source elements in the AluYh3a1 subfamily, however, the results of the simulations of AluYh3a1 in humans suggest that secondary source elements may have contributed to its proliferation. It is, of course, possible that this subfamily contains several source elements all of which are identical to the subfamily consensus. These would not be identified by examining patterns of shared mutations. In humans, the optimal number of source elements is relatively high, at 8.08, however, the most frequently observed number of sources in successful runs, in 277 out of 1418 (19.5%) runs, is 1, making master gene proliferation a possibility for this subfamily (Table 3.3). The most successful parameter combination, pA = 0.7, pT = 4x10^{-7}, results in 45% of runs within the required range, with a very high average number of sources of 29.9. The total range of source elements observed in successful runs under condition 1 is between 1 and 52 (Figure 3.7). Under condition 2 (Table 3.4), the number of sources is much smaller, as the number of shared mutations in this subfamily is relatively small. The most frequent value is still 1, but its frequency increases to 45 out of 129, or 34.9%, under condition 2. The most successful combination, with 12% success, has a small pA of 0.04, coupled with a pT of 4x10^{-6}, compared with the relatively high pA value of 0.7 under condition 1. The most successful number of sources therefore drops to 4.01. The average number of sources in successful runs is also lower, at 3.26, and the total possible range of sources is much more restricted, between 1 and 12 (Figure 3.8). However, under condition 3 (Table 3.5), numbers of sources are again high. The total possible range is once again between 1 and 52, and although one source is still the most frequent, its frequency drops to 7.61% (Figure 3.9). The average number of sources increases above that of condition 1, to 16.81. There are two equally successful parameter combinations under condition 3, each with 14% of runs falling within the required range. These combinations, pA = 0.3 with pT = 8x10^{-7}, and pA = 0.4 with pT = 7x10^{-7}, generate an average of 17.84 and 22.06 sources, respectively.



Source	Frequency
1	2844
2	277
3	66
4	26
5	12
6	4
7	4
8	2
9	2
10	4
11	0
12	2
13	1
14	0
15	0
16	1
17	0
18	0
19	0
20	1

Figure 3.10: The frequency of each possible number of source elements in successful runs for the AluYh3a1 subfamily in chimpanzees, under condition 1.

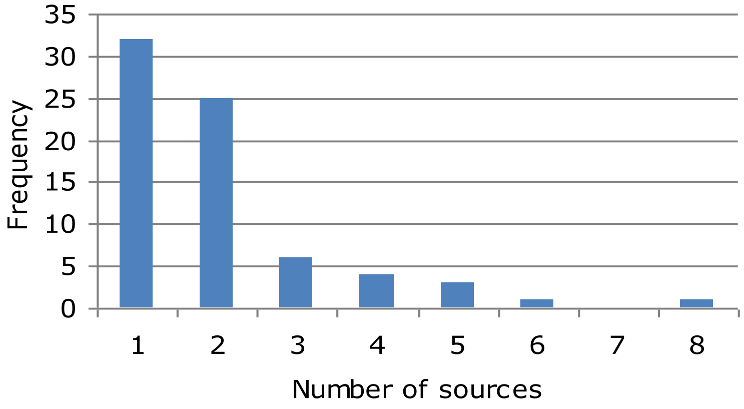


Figure 3.11: The frequency of each possible number of source elements in successful runs for the AluYh3a1 subfamily in chimpanzees, under condition 2.

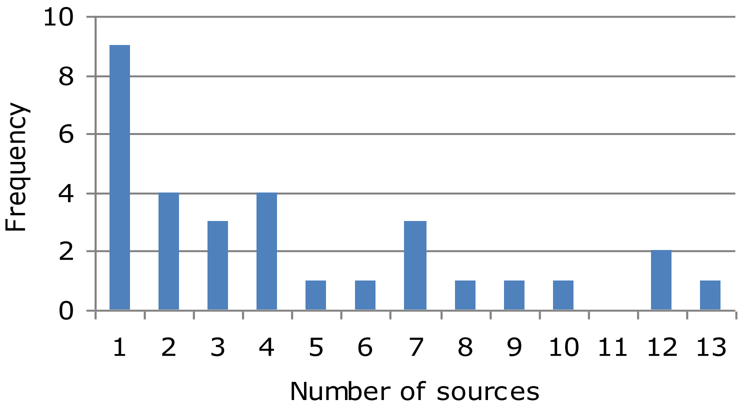


Figure 3.12: The frequency of each possible number of source elements in successful runs for the AluYh3a1 subfamily in chimpanzees, under condition 3.

The estimates of numbers of source elements in the Yh3a1 family in chimpanzees is lower, and may suggest that human-specific source elements became active following the divergence of humans and chimpanzees. Under condition 1, the most frequently observed number of sources is again 1 (Table 3.3), but its frequency is much higher for this species, in 2844 out of 3246 (87.6%) successful runs. The most successful parameter combination is radically different in the chimpanzee compared with the human, which had a very high value of pA, with pA equal to 0, and pT equal to 5×10^{-6}, such that the number of sources in all runs is 1. This combination generated 78 runs out of 100 within the required range. The average number of sources is also lower, at 1.22 in the chimpanzee compared with 12.68 in the human, and the range of possible sources is more restricted, falling between 1 and 20 (Figure 3.10). The estimate of the number of source elements in the chimpanzee increases under condition 2 (Figure 3.11), and further under condition 3 (Figure 3.12), reaching an average of 4.45 under the latter.

3.3.2.4 AluYi6

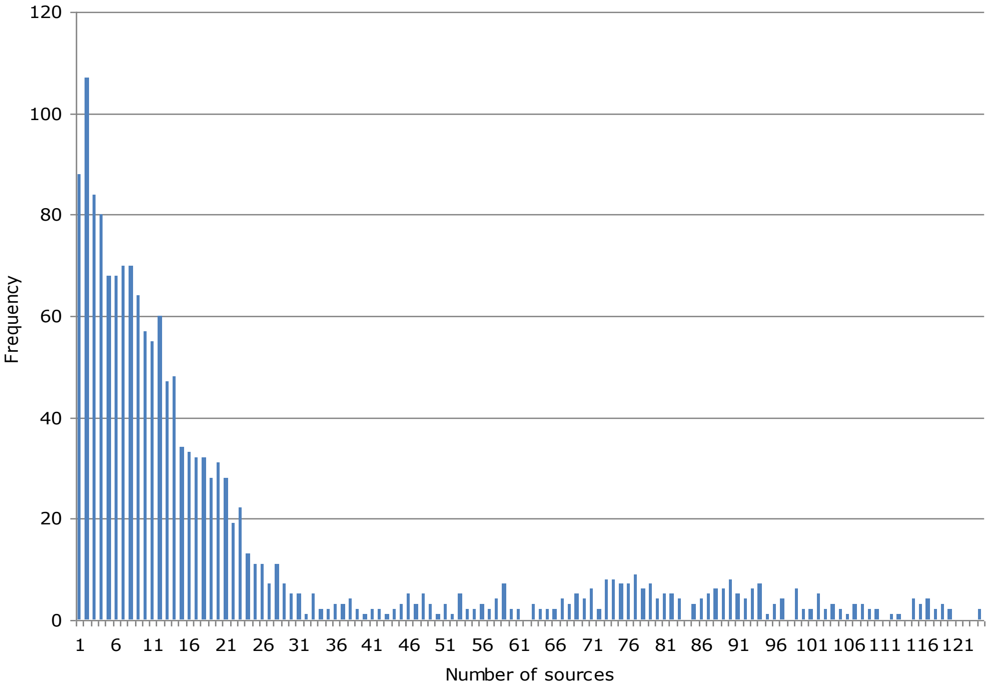


Figure 3.13: The frequency of each possible number of source elements in successful runs for the AluYi6 subfamily in humans, under condition 1.

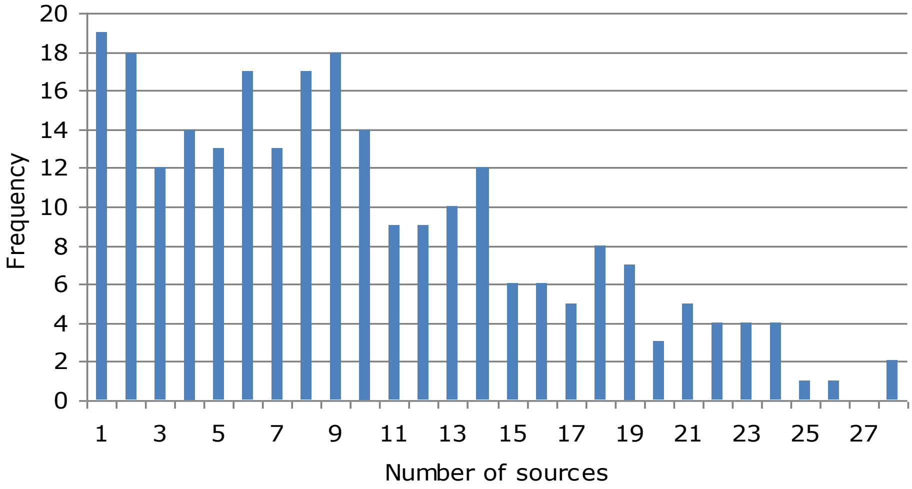


Figure 3.14: The frequency of each possible number of source elements in successful runs for the AluYi6 subfamily in humans, under condition 2.

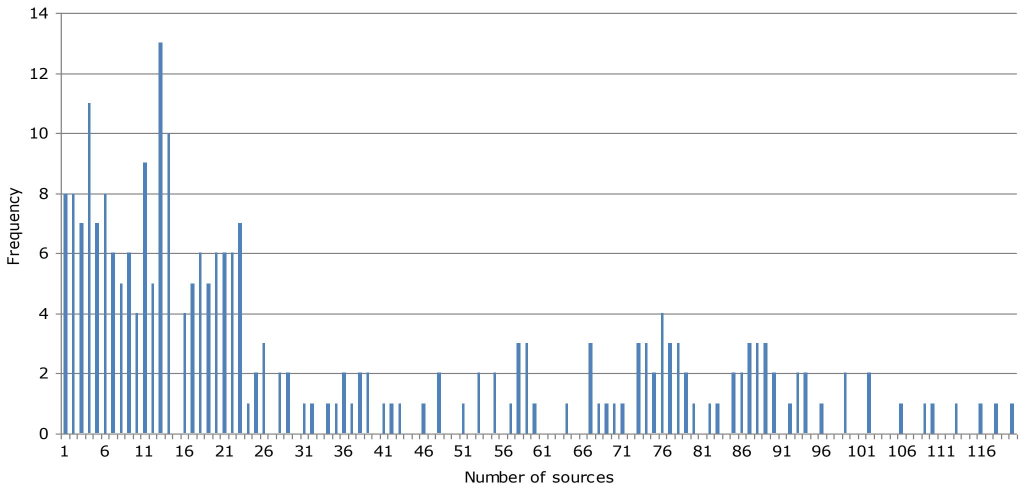
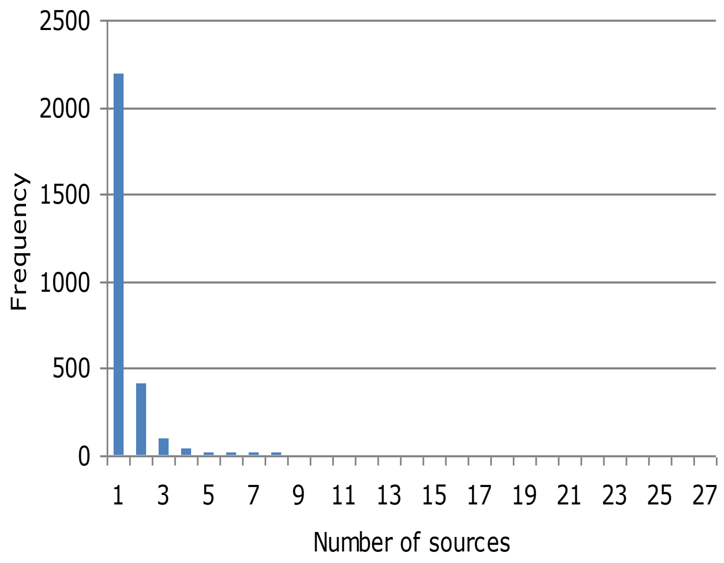


Figure 3.15: The frequency of each possible number of source elements in successful runs for the AluYi6 subfamily in humans, under condition 3.

The optimal number of sources for the AluYi6 subfamily estimated by the program in both humans and chimpanzees is in good agreement with the estimates from polymorphism data and patterns of shared mutations, at 10.29 and 4.52, respectively. The total range of number of sources capable of generating the AluYi6 family is extremely large in humans, ranging between 1 and 124 (Figure 3.13) under condition 1. The most successful average number of sources is 11.47 (Table 3.3), similar to the optimal value, generated from the parameter combination pA = 0.07, pT = 5x10^{-6}, with 37% success. Both one source element and two source elements are observed equally frequently in successful runs, in 41 out of 1598 (2.57%) of runs. However, although these are the most frequent, the vast majority of successful runs require many more sources. The average of all successful runs is 22.74 sources. Under condition 2 (Table 3.4), the most frequent is a single source, in 7.57% of successful runs. When shared mutations are taken into account, the number of source elements estimated by the program decreases, with the total possible range restricted between 1 and 28 (Figure

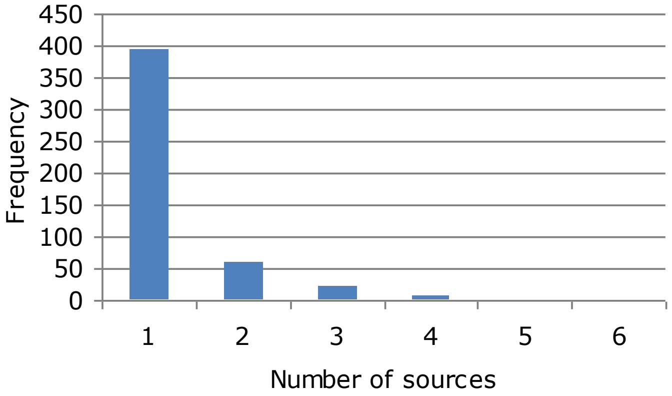
3.14), and the average decreasing to 9.74. The most successful parameter combination, with 12% success, is pA = 0.05, with pT = 7x10^{-6}, which results in an average of 8.72 source elements. When pi is taken into consideration under condition 3 (Table 3.5), the estimate of the number of source elements once again increases, with 13 being the most commonly observed number in 5.10% of successful runs. The most successful combination possesses a relatively high value of pA, 0.4, with pT of 1x10^{-6}, resulting in a large number of sources averaging at 47.06. The average across all successful runs increases to 34.30, and the total possible range once again expands, covering a wide range of numbers of source elements between 1 and 120 (Figure 3.15).

The AluYi6 family in chimpanzees yields interesting results, as it is the only subfamily investigated for which successful runs were obtained within the appropriate range for all four test statistics: theta, pi, time and shared mutations. This might indicate that the evolution of this subfamily in reality fits well with the assumptions of the program, such as no gene conversion, equal retrotransposition rates of different source elements, and a constant retrotransposition rate over time. A much smaller number of sources is estimated for the AluYi6 family in chimpanzees compared with humans, partly due to its lower copy number. The most frequent number of sources under all conditions was 1, which was extremely common, found in 78.97% of successful runs under condition 1, and 81.91% and 48.81% under conditions 2 and 3, respectively. A single source element is also most frequent in those runs which are successful for all four test statistics (condition 4), where 97.62% of successful runs possess a single source element. As for other subfamilies, the average number of sources is highest under condition 3 (Table 3.5), at 2.35, but is small, at 1.02 under condition 4, 1.26 under condition 2 (Table 3.4) and 1.38 under condition 1 (Table 3.3). The most successful parameter combination under condition 1 (pA =7x10^{-4}, pT =1x10^{-5}) yields 79% of runs within the required range, and an average number of sources of 1.03. Under conditions 2 and 4, the same parameter combination, pA = 9x10^{-4} with pT = 9x10^{-6}, is most successful, with an average of 1.1 sources and 14% and 7% of runs successful, respectively. Under condition 3, a different parameter combination, with a much larger value of pA, 0.07, with pT = 3x10^{-6} is most successful, with 7% success, yielding a higher average number of sources of 5.8. The total range of sources capable of generating the AluYi6 subfamily in chimpanzees is between 1 and 20 under condition 1 (Figure 3.16), restricted to 1 – 17 under condition 3 (Figure 3.18), and 1 – 6 under condition 2 (Figure 3.17). Under condition 4, only runs which possessed either 1 or 2 sources were capable of generating successful subfamilies.



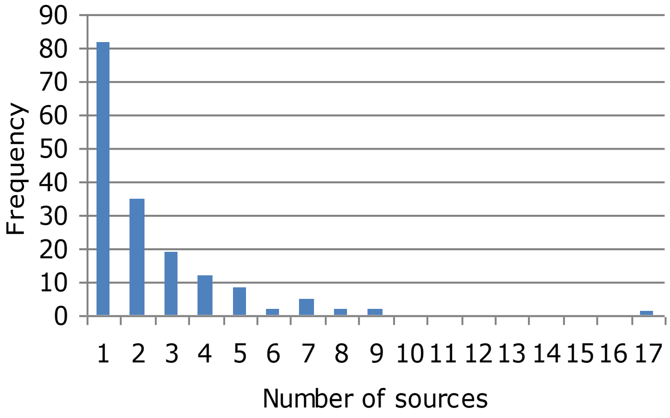
Source	Frequency
1	2200
2	406
3	93
4	39
5	17
6	8
7	7
8	6
9	2
10	1
11	1
12	0
13	0
14	0
15	0
16	1
17	2
18	0
19	1
20	0
21	0
22	0
23	1
24	0
25	0
26	0
27	1

Figure 3.16: The frequency of each possible number of source elements in successful runs for the AluYi6 subfamily in chimpanzees, under condition 1.



Source	Frequency
1	394
2	59
3	21
4	5
5	1
6	1

Figure 3.17: The frequency of each possible number of source elements in successful runs for the AluYi6 subfamily in chimpanzees, under condition 2.



Source	Frequency
1	82
2	35
3	19
4	12
5	8
6	2
7	5
8	2
9	2
10	0
11	0
12	0
13	0
14	0
15	0
16	0
17	1

Figure 3.18: The frequency of each possible number of source elements in successful runs for the AluYi6 subfamily in chimpanzees, under condition 3.

For all subfamilies, under condition 1, it was possible to generate subfamilies resembling those observed in reality under the master gene model. However, it was only possible to generate subfamilies for which test statistics fell within the required range under the transposon model, whereby all elements are active (pA = 1), for the AluYh7 subfamily. This subfamily has only been retrotransposing in the recent past, and it is likely that in these simulations where pA = 1, few mutations are contained within the source elements. If shared mutations are taken into consideration, the transposon model can also be rejected for AluYh7, with the number of shared mutations required under these conditions (25 – 48) much higher than that observed in reality. From analysis of AluYh3a1 and AluYi6, which are present in both humans and chimpanzees, it appears that the number of source elements for each of these families is higher in humans than in chimpanzees. This may be a consequence of the elevated retrotransposition rate in humans, such that a higher copy number is achieved and therefore a larger number of elements in humans operate as sources. For AluYh3a1, however, the copy number in humans and chimps is not very different. It may be that more elements are capable of retrotransposing in humans, and, as a consequence, the retrotransposition rate is higher. Estimates from the program are in line with estimates from both the sequence data and polymorphism, and inclusion of shared mutations as a condition for success appears to improve this relationship, although rates of success per parameter combination are relatively low. This may suggest that there are many factors influencing the evolution of Alu elements which are unknown, or do not fit with the assumptions of the program. The total range of values for number of sources which are able to generate successful runs, and therefore may apply to the real subfamilies, are large in many cases, such as human AluYi6. The results do however rule out many scenarios for the evolution of these subfamilies, and give an indication of the number of source elements which are most likely to have operated in the proliferation of these subfamilies.

Subfamily	Sources		
	Sequence	Polymorphism	Program (optimal)
Yg6	3 – 4	6	3.59
Yh7	1 – 2	2	1.54
Yh3a1 (human)	1	-	8.08
Yh3a1 (chimp)	1	-	6.55
Yi6 (human)	6 – 14	8	10.29
Yi6 (chimp)	4 – 7	3	4.52

Table 3.6: The number of source elements for each subfamily, estimated using sequence data, polymorphism data, and the program. Levels of polymorphism were not assessed for AluYh3a1, as discussed in chapter 2. The optimal number of sources from the program corresponds to the average number of sources for the parameter combination which gave the best match, on average, to the test statistics theta, shared mutations and time for each subfamily.

3.3.3 Rate of Retrotransposition

Subfamily	Optimal retrotransposition rate (pT per source per year)
AluYh7	$7x10^{-6}$
AluYg6	$4x10^{-5}$
AluYh3a1 (human)	$3x10^{-6}$
AluYh3a1 (chimp)	$2x10^{-6}$
AluYi6 (human)	$8x10^{-6}$
AluYi6 (chimp)	$5x10^{-5}$

Table 3.7: The optimal rate of retrotransposition for each subfamily. This is equivalent to the value of the parameter pT which, in combination with a particular value of pA, resulted in average values for the testing parameters proliferation time, theta and shared mutations (i.e. condition 2) which most closely resembled the real data.

The optimal rate of retrotransposition per source element per year corresponds to the value of the parameter pT that gave, as an average across 100 runs, the best match to the values for theta, proliferation time and shared mutations. This value was found to vary between subfamilies, in some cases by more than an order of magnitude (Table 3.7). The retrotransposition rate was also found to vary between species, where a subfamily was present in both humans and chimpanzees, in the case of AluYi6. However, for AluYh3a1, which is also found in both species, the optimal retrotransposition rates only differ very slightly. This could be indicative of a similar number of source elements for this subfamily in humans and chimpanzees, as the retrotransposition rate is given per source per year. This would be unsurprising, given that the majority of elements in this subfamily are shared by both species. The estimates of numbers of sources above, however, suggest a greater number of sources in humans. For AluYi6, there appear to be many secondary source elements operating in humans, but most of these seem to have very low activity, so the average rate per source is low. In chimps, however, there are few sources, but which seem to have relatively high activity, such as the chimp-specific secondary source gene which has generated approximately one third of the chimpanzee elements discussed in chapter 2. The assumption that each source gene has the same level of activity therefore appears to be unrealistic in at least some cases. The fact that different subfamilies within a species can have different rates of retrotransposition suggests that this is also likely to be true of different source elements within a subfamily. Table 3.7 shows the optimal value of pT for each subfamily, that is, the value which generates simulated subfamilies most closely resembling the real data. As for numbers of sources, discussed above, the most successful, most frequent, average and total possible range of values for pT are given for each subfamily under the three conditions described previously. Although each value of pT within the total possible range for each subfamily is capable of generating results resembling the real data, some values are more likely than others, as many demonstrate a very low percentage of successful runs.

Subfamily	Retrotransposition rate per source per year			
	Most successful	Most frequent	Average	Total possible
AluYh7	$6x10^{-6}$	$6x10^{-6}$	$5.5x10^{-6}$	$4x10^{-7}$ – $1x10^{-5}$
AluYg6	$1x10^{-4}$	$1x10^{-4}$	$7.98x10^{-5}$	$8x10^{-7}$ – $9x10^{-4}$
AluYh3a1 (human)	$4x10^{-7}$	$4x10^{-6}$	$3.3x10^{-6}$	$3x10^{-7}$ – $1x10^{-5}$
AluYh3a1 (chimp)	$5x10^{-6}$	$5x10^{-6}$	$5.3x10^{-6}$	$4x10^{-7}$ – $9x10^{-6}$
AluYi6 (human)	$5x10^{-6}$	$5x10^{-6}$	$5.6x10^{-6}$	$4x10^{-7}$ – $3x10^{-5}$
AluYi6 (chimp)	$1x10^{-5}$	$1x10^{-5}$	$8.6x10^{-6}$	$9x10^{-7}$– $1x10^{-5}$

Table 3.8: Retrotransposition rates, pT, per source element per year required to generate subfamilies resembling those observed in reality estimated by the program under condition 1.

Subfamily	Retrotransposition rate per source per year			
	Most successful	Most frequent	Average	Total possible
AluYh7	$6x10^{-6}$	$5x10^{-6}$	$6.3x10^{-6}$	$2x10^{-6}$ – $1x10^{-5}$
AluYg6	$1x10^{-5}$	$1x10^{-5}$	$2.02x10^{-5}$	$5x10^{-6}$ – $6x10^{-5}$
AluYh3a1 (human)	$4x10^{-6}$	$4x10^{-6}$	$5.1x10^{-6}$	$2x10^{-6}$ – $1x10^{-5}$
AluYh3a1 (chimp)	$3x10^{-6}$	$3x10^{-6}$	$3.3x10^{-6}$	$2x10^{-6}$ – $5x10^{-6}$
AluYi6 (human)	$7x10^{-6}$	$5x10^{-6}$	$6.7x10^{-6}$	$3x10^{-6}$ – $1x10^{-5}$
AluYi6 (chimp)	$9x10^{-6}$	$9x10^{-6}$	$8.5x10^{-6}$	$4x10^{-6}$ – $1x10^{-5}$

Table 3.9: Retrotransposition rates, pT, per source element per year required to generate subfamilies resembling those observed in reality estimated by the program under condition 2. Unlike in table 3.7, where optimal values are shown, these values of pT correspond to the values which, in conjuction with particular values for pA, resulted in the greatest number of successful runs (most successful), was most frequently observed in successful runs (most frequent), was the average pT value across successful runs (average), along with the total range of values for pT which were able to generate at least one successful run (total possible).

Subfamily	Retrotransposition rate per source per year			
	Most successful	Most frequent	Average	Total possible
AluYh7	-	-	-	-
AluYg6	$1x10^{-5}$	$1x10^{-5}$	$1.01x10^{-5}$	$2x10^{-6}$ – $4x10^{-5}$
AluYh3a1 (human)	$7x10^{-7}$	$7x10^{-7}$	$1.4x10^{-6}$	$3x10^{-7}$ – $6x10^{-6}$
AluYh3a1 (chimp)	$3x10^{-6}$, $8x10^{-7}$	$3x10^{-6}$	$2.2x10^{-6}$	$6x10^{-7}$ – $3x10^{-6}$
AluYi6 (human)	$1x10^{-6}$	$1x10^{-6}$	$2.8x10^{-6}$	$5x10^{-7}$ – $1x10^{-5}$
AluYi6 (chimp)	$3x10^{-6}$	$5x10^{-6}$	$5.7x10^{-6}$	$1x10^{-6}$ – $1x10^{-5}$

Table 3.10: Retrotransposition rates, pT, per source element per year required to generate subfamilies resembling those observed in reality estimated by the program under condition 3. No successful runs for the AluYh7 family possessed a pi value within the required range.

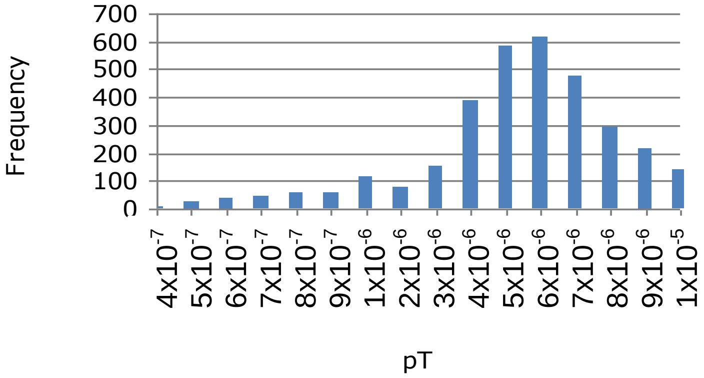


Figure 3.19: Frequency of the total range of values for pT which are able to generate successful runs for the AluYh7 family under condition 1.

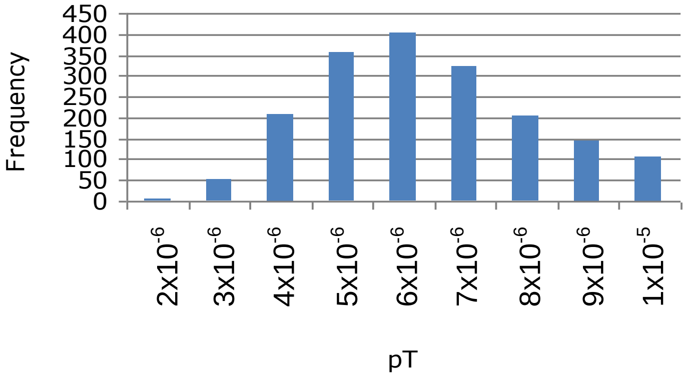


Figure 3.20: Frequency of the total range of values for pT which are able to generate successful runs for the AluYh7 family under condition 2.

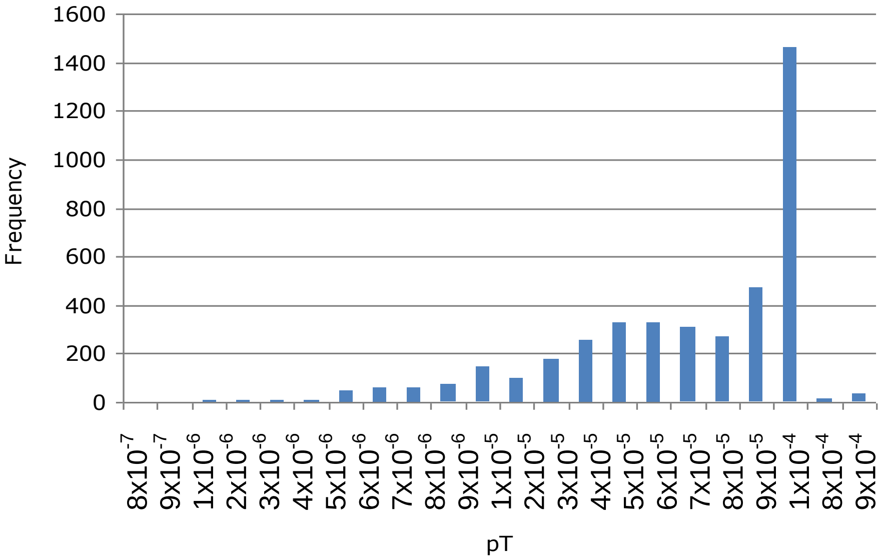


Figure 3.21: Frequency of the total range of values for pT which are able to generate successful runs for the AluYg6 family under condition 1. The large spike at pT =1x10^{-4} followed by no instances of pT = 2x10^{-4} is a consequence of both using a logarithmic scale to increment values of pT, and also the required association with particular values of pA in order to generate a successful run.

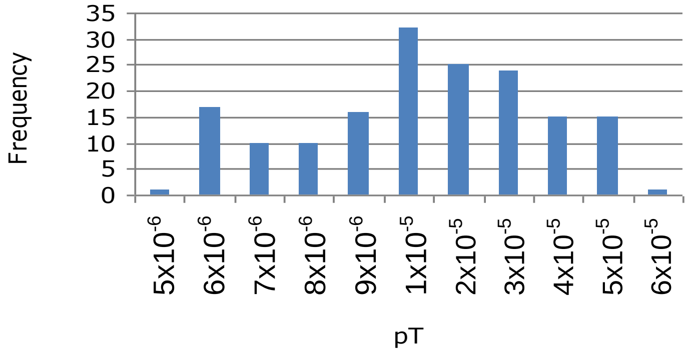


Figure 3.22: Frequency of the total range of values for pT which are able to generate successful runs for the AluYg6 family under condition 2.

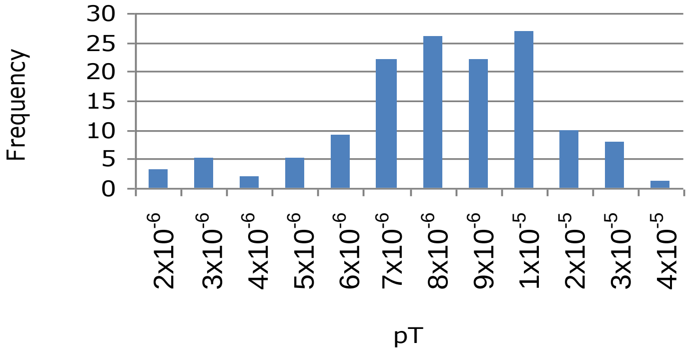


Figure 3.23: Frequency of the total range of values for pT which are able to generate successful runs for the AluYg6 family under condition 3.

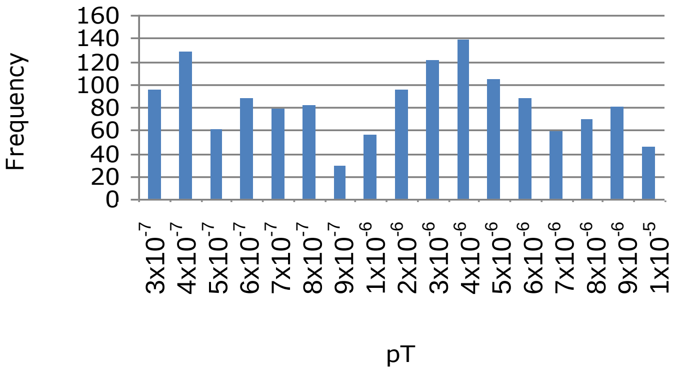


Figure 3.24: Frequency of the total range of values for pT which are able to generate successful runs for the AluYh3a1 family in humans under condition 1.

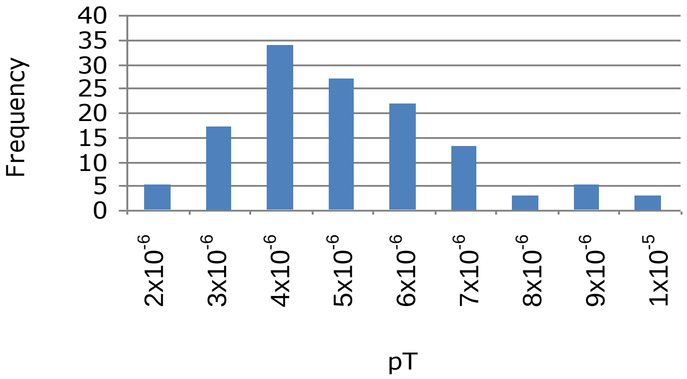


Figure 3.25: Frequency of the total range of values for pT which are able to generate successful runs for the AluYh3a1 family in humans under condition 2.

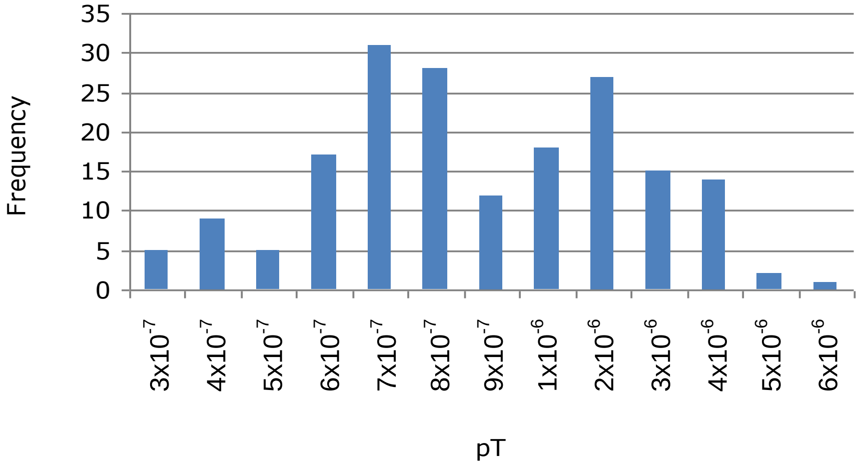


Figure 3.26: Frequency of the total range of values for pT which are able to generate successful runs for the AluYh3a1 family in humans under condition 3.

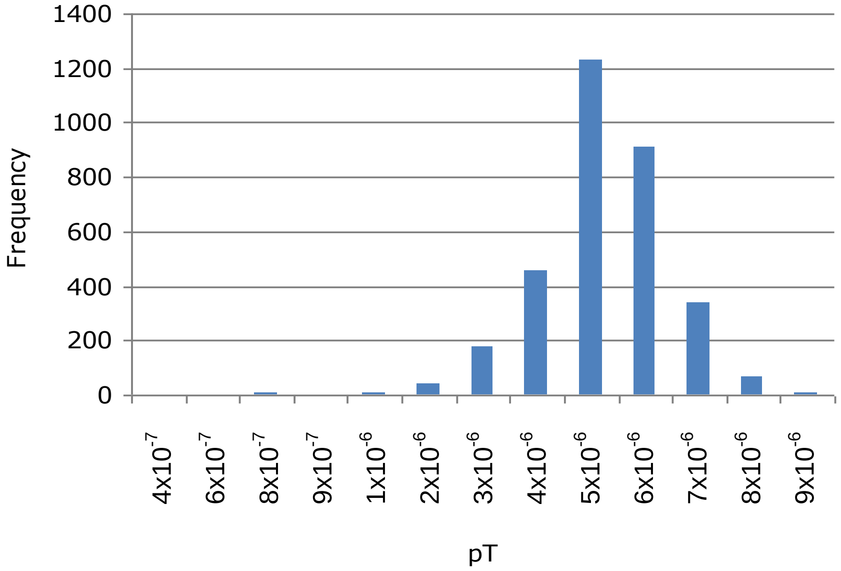


Figure 3.27: Frequency of the total range of values for pT which are able to generate successful runs for the AluYh3a1 family in chimpanzees under condition 1.

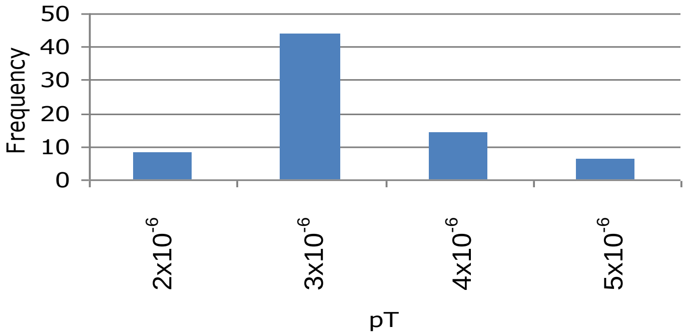


Figure 3.28: Frequency of the total range of values for pT which are able to generate successful runs for the AluYh3a1 family in chimpanzees under condition 2.

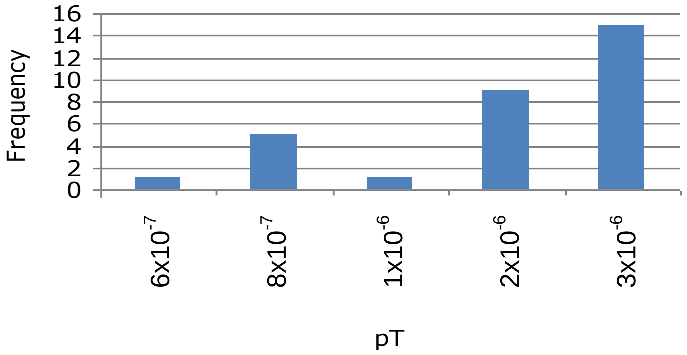


Figure 3.29: Frequency of the total range of values for pT which are able to generate successful runs for the AluYh3a1 family in chimpanzees under condition 3. Successful runs are not generated for pT = 4x10^{-6}, however, it is likely that if values greater than 3x10^{-6} by a smaller increment were tested, success would be observed.

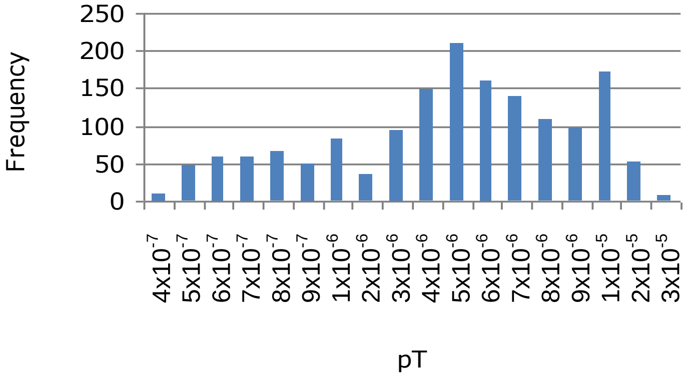


Figure 3.30: Frequency of the total range of values for pT which are able to generate successful runs for the AluYi6 family in humans under condition 1.

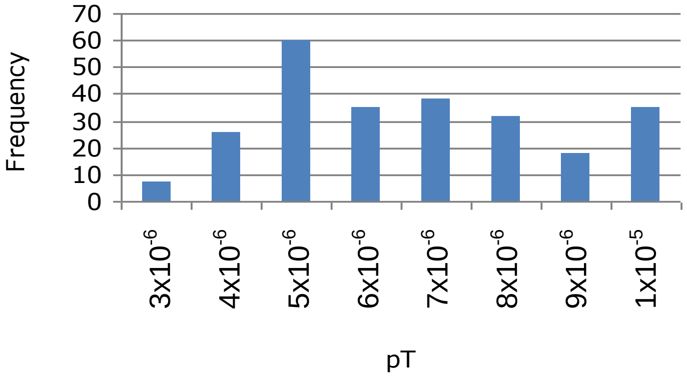


Figure 3.31: Frequency of the total range of values for pT which are able to generate successful runs for the AluYi6 family in humans under condition 2.

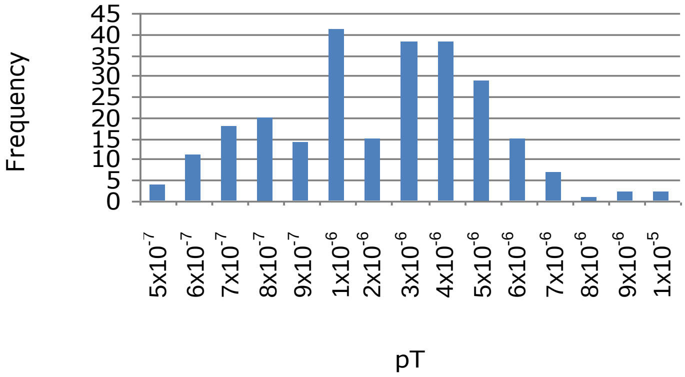


Figure 3.32: Frequency of the total range of values for pT which are able to generate successful runs for the AluYi6 family in humans under condition 3.

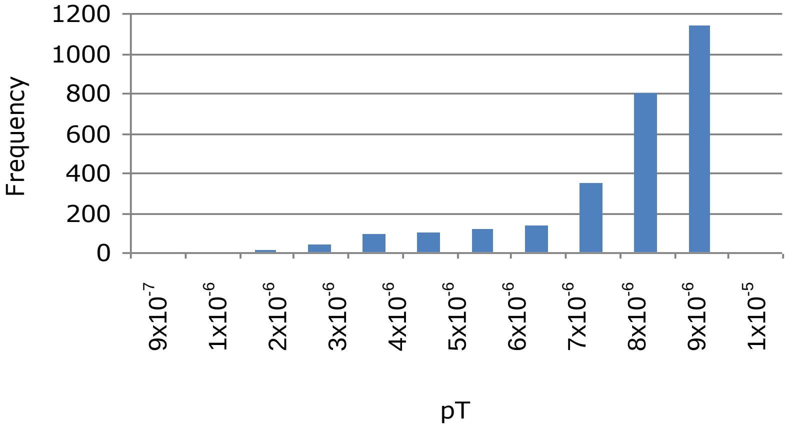


Figure 3.33: Frequency of the total range of values for pT which are able to generate successful runs for the AluYi6 family in chimpanzees under condition 1. pT = 2x10^{-5} did not generate successful runs.

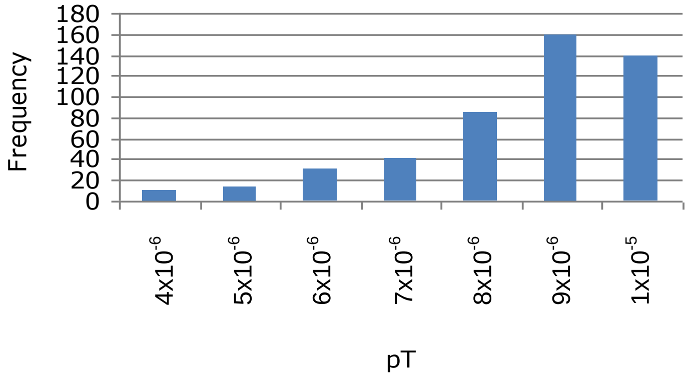


Figure 3.34: Frequency of the total range of values for pT which are able to generate successful runs for the AluYi6 family in chimpanzees under condition 2.

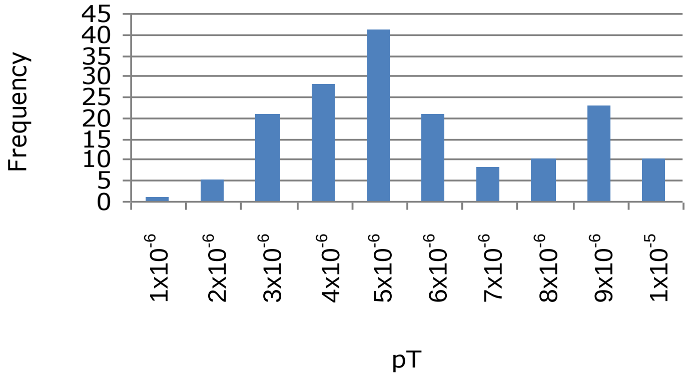


Figure 3.35: Frequency of the total range of values for pT which are able to generate successful runs for the AluYi6 family in chimpanzees under condition 3.

There is more consistency between the three conditions and the five estimates for each subfamily under each condition for retrotransposition rate than for number of sources (Tables 3.8-3.10). For AluYg6, the total possible range extends to quite a large value of pT (Figures 3.22-3.23), which is required to generate such a large subfamily in the given time where the number of active elements available is low. Retrotransposition rates do appear to vary between different subfamilies, which suggests that variation between source elements within a subfamily is also likely to exist. For the subfamilies which are found in both humans and chimpanzees, it is interesting to compare the estimated retrotransposition rates between the two species. Generally, the retrotransposition rate per source per year is much smaller for the AluYh3a1 subfamily in humans (Figures 3.24-3.26) compared with chimpanzees (Figures 3.27-3.29), as a consequence of an estimation of a larger number of sources in humans. The AluYh3a1 subfamily in chimpanzees has reached a slightly lower copy number, which appears to have been brought about by more active, but fewer, source elements. For the AluYi6 subfamily, again a larger number of sources is estimated by the program, and consequently smaller values of pT (Figures 3.30-3.32) are required to generate the subfamily in the required time, despite a substantially higher copy number of AluYi6 in humans. It does indeed appear that the increased retrotransposition rate in humans may be due to an increase in the number of elements which are able to effectively retrotranspose, rather than an increase in the activity of each element. However, it should be borne in mind that these results do suggest variation in the activity of individual source elements in young Alu subfamilies, which are assumed to have the same level of activity in the program.

3.4 Conclusions

The program presented here enables the estimation of the values for parameters affecting the evolution of young Alu subfamilies, such as the rate of retrotransposition and number of source elements, using genomic sequence data. The simulations suggest that the master gene model is not the most likely model to explain the proliferation of many young Alu subfamilies, although it is a possible candidate under most conditions. Most subfamilies seem to require the activity of several, and sometimes many, as in the case of human AluYi6, secondary source elements. The transposon model was unable to generate subfamilies resembling those observed in reality, except in the case of AluYh7 under condition 1. The program makes some assumptions about Alu evolution, some of which have been shown to be unlikely, such as each source element having a constant rate of retrotransposition, and each source element in a subfamily having the same level of activity. The results of the simulations suggest, at least in the case of AluYh3a1, that the rate of retrotransposition may have varied throughout evolutionary history. However, the program provides an appropriate model for the evolution of young Alu subfamilies, as the results obtained for each subfamily are reasonable based on available data, such as sequence analysis and presence or absence polymorphism.

Chapter 4 - Horizontal transfer of transposable elements in *Drosophila*

4.1 Introduction

Horizontal transfer is the process by which genetic material is transferred from one species to another, for example due to introgression, or through a vector such as a virus or parasite. This phenomenon is particularly interesting to investigate in *Drosophila*, as these species rapidly eliminate transposable elements from their genome, such that transfer to other species might provide the only means of escape and survival for these elements (Loreto et al. 2008). Horizontal transfer of several families of transposable element has been shown to have occurred between members of the *Drosophila* genus. A review of all reported cases of horizontal transfer has recently become available, and indicates that horizontal transfer in *Drosophila* is most frequent for the DNA transposons, with slightly fewer instances for the LTR retrotransposons. Horizontal transfer of the non-LTR retrotransposons was shown to be very rare (Loreto et al. 2008).

Before horizontal transfer can be inferred to have occurred, there are some basic requirements that must be met (reviewed by Loreto et al. 2008). Firstly, there must be geographical, temporal and ecological overlap between the donor and recipient species. The two species involved in the putative transfer event must be found in the same, or overlapping, ecological niches, in the same geographical location at the same time. Secondly, there must be an appropriate vector available that would be able to transfer the transposable element from one species to the other. Investigating potential vectors is beyond the range of this study, and it will be assumed that, when horizontal transfer is inferred to have occurred, an appropriate vector would be available.

For example, it can be assumed that introgression, intracellular parasites such as *Wolbachia*, and extracellular parasites, are likely routes of transfer for any combination of *Drosophila* species. Therefore, the only prior requirement that will be investigated is that the two species overlap. To test the validity of the methods employed in this study, which are described below, geographical distribution of the species will not be examined until putative cases of horizontal transfer have been inferred. That is, transposable elements of the same family in different species will be compared regardless of whether or not the host species overlap. If the methods are appropriate, no cases of horizontal transfer should be inferred between species which do not overlap geographically.

Horizontal transfer of transposable elements is believed to occur more frequently among the *Drosophila* species than in many other eukaryotic groups that have been investigated (Loreto et al. 2008). It may be that elements have more opportunity to undergo transfer, perhaps as a result of frequent contact with vectors. However, it is argued that horizontal transfer may be an essential step in the lifecycle of many of the transposable element families which exist in the *Drosophila* genomes (Loreto et al. 2008). This is because transposable elements are rapidly eliminated from *Drosophila* genomes, to the extent that if two extremely closely-related species, such as *D. pseudoobscura* and *D. persimilis*, which diverged around two million years ago, are examined, in the majority of cases it would be expected that no orthologous elements would be found. For this reason, it may be that were transposable elements not to undergo horizontal transfer in *Drosophila* species, infecting naïve genomes that are unable to control their proliferation, the vast majority of them would be eliminated and therefore would not be observed in the contemporary genome sequences of these species.

Horizontal transfer of transposable elements between *Drosophila* species might be suspected if one or more of the following observations are made (reviewed by Loreto et al. 2008). Firstly, if the identity between elements from different species is higher than the identity between the host genes in those species, this would suggest that the transposable elements share more recent common ancestry than the genes, and therefore the species, themselves. This line of evidence to support horizontal transfer makes the assumption that the host genes are more highly constrained than the transposable elements. Many transposable elements do contain open reading frames and recognition sequences for enzymatic activity which can be assumed to be constrained in the sense that the sequence would not be capable of transposition were these sequences to mutate to a great extent. However, once a transposable element has integrated, there is not any direct selection on it to retain its function (Bergman and Bensasson 2007). Constraint may be observed however in, for example, a population of recently integrated elements, which must have all been functional at the moment of integration, and therefore may be to some extent constrained. However, in this study, divergence will be compared with that of only the coding region of a host gene, which is presumably under greater constraint than the transposable element sequences. Constraint does however need to be borne in mind as a confounding factor when looking at divergence. The strength of this line of evidence needs also to be considered. Although constraint may account for similarity between elements in different species, cases where divergence between a pair or group of elements in different species is strikingly small, among an overall population of elements which shows considerable variation between species, would be strong evidence that the high identity was due to recent common ancestry rather than constraint.

Secondly, phylogenetic trees can be constructed using the transposable element sequences of a particular family from all the species in which it is present. If there is incongruence between the topology of the tree of the elements and the known host species phylogeny, such that the relationships inferred by the tree of the transposable element sequences are not consistent with the known relationships between the host *Drosophila* species, this also supports the hypothesis of horizontal transfer (Figure 4.1). Examination of incongruent phylogenies, as well as supporting the case for horizontal transfer, can also be useful in determining the potential direction of any such transfer. For example, had horizontal transfer introduced a transposable element family from a donor species into a recipient species in which the family was previously absent, all elements in the recipient species should cluster together on the tree, within the clade of elements from the donor species. This would allow both the donor and recipient species to be identified, and therefore the direction of transfer inferred. This is only possible where the family is present in greater than two host species.

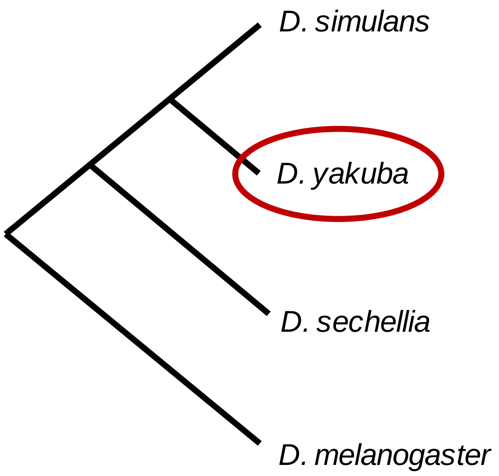


Figure 4.1: Phylogenetic incongruence observed on a tree constructed using transposable element sequences. In this case, these relationships suggest horizontal transfer has occurred from *D. simulans* into *D. yakuba*.

However, there are certain caveats associated with this line of evidence for horizontal transfer, and alone, it is not particularly convincing. Firstly, it is possible for host and transposable element trees to be congruent even if horizontal transfer has occurred. For example, if a transposable element family is present in a single species, and is then introduced from that species into a naïve genome, all of the elements in the second species would be each other's closest relatives, as they would all be descended from the element involved in the transfer. Were it possible to root the tree, this clade of elements from the recipient species should fall within the clade of elements from the donor species, however, in the absence of a means to root the tree, the phylogeny would appear congruent and horizontal transfer could not be inferred. Congruent phylogenies can also be generated following horizontal transfer events where a greater number of species are involved, depending on the relationships between these hosts. For example, transfer from the ancestor of a group of species into the closest relative outside of those species is likely to generate a congruent phylogeny. For example, were transfer to have occurred from the ancestor of *D. melanogaster*, *D. simulans* and *D. sechellia* into *D. yakuba*, the elements from *D. simulans* and *D. sechellia* would be most closely related to each other, and then to the elements from *D. melanogaster*, and then those from *D. yakuba*. This is consistent with the relationships between the host species. More generally, where ancient transfers are involved, such as between the Sophophora and Drosophila subgenera prior to their diversification, congruent phylogenies with respect to these relationships are a likely possibility. Transfers which result in the introduction of a transposable element family into a species in which it has never been, or is no longer, present, are more likely to generate congruent phylogenies than transfers among species in which a population of elements belonging to that family is already established. In these cases, provided copy number of elements in individual species exceeds one, it is hopeful that phylogenetic incongruence would be observed. Elements in the recipient species which are descended from the transfer event should be found within a clade of elements from the donor species, whereas other elements from this species should be found elsewhere on the phylogeny, in a position congruent with the relationship between those two host species.

The reliability of the observation of phylogenetic incongruence itself is also an issue, as it is possible for phylogenies produced using transposable element sequences and those produced using host genes to be incongruent even if horizontal transfer has not occurred. It is for this reason that phylogenetic incongruence alone does not provide convincing evidence to support the hypothesis of horizontal transfer. The primary explanation for phylogenetic incongruence in the absence of horizontal transfer is differential retention of ancestral polymorphism (Figure 4.2). This is not polymorphism in the more typical sense of allelic variation at a single locus, but rather variation in sequence between elements within a family that are located in different loci. Ancestral polymorphism is particularly a problem when trying to resolve relationships between transposable elements in pairs of host species which are extremely closely related, such as *D. simulans* and *D. sechellia*, or *D. persimilis* and *D. pseudoobscura*. A variable population of transposable elements of a particular family can be inferred to have been present in the common ancestor of these species, which may have been accumulating mutations for some time. If descendents of many of these variable elements are present in the host species following their divergence, which is most likely for the most closely-related species for which divergence time is small, these may form different groups on a phylogeny, much in the same way as paralogous genes. Ancestral polymorphism can also generate incongruent relationships on a phylogeny across greater divergence times. Assuming a variable population of elements was present in the common ancestor, over time, certain variants may become fixed in certain lineages, incorrectly generating the impression of transfer between species.

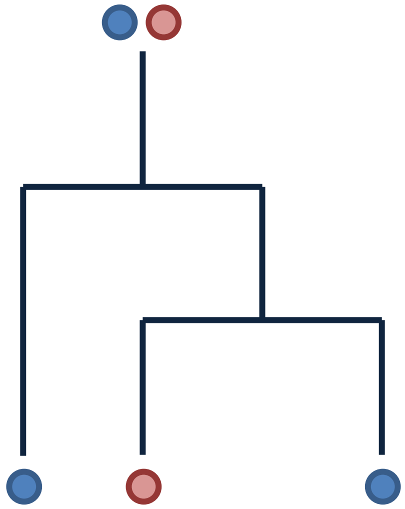


Figure 4.2: Incomplete lineage sorting of an ancestral polymorphism. Two "subfamilies" of a particular transposable element family are present in the ancestor, with each of the descendents having inherited only one.

Thirdly, the distribution of the transposable element family across the host *Drosophila* species can be examined, that is, in which species the family is present, and in which it is absent. If the distribution is “patchy”, for example if the family is unexpectedly absent from particular lineages in spite of its presence in closely related species, this provides some support to the hypothesis of horizontal transfer (Figure 4.3). For example, a family may have been inherited from a common ancestor and be found in all of the descendents of that ancestor, as well as in a single very distantly related species, but not in the closest relatives of that distant relative. Such a distribution may suggest that horizontal transfer has introduced the transposable element family into the distantly related species from a member of the other group. However, it is this line of evidence that perhaps needs to be treated with the most caution. As described previously, deletion of transposable element sequences from *Drosophila* genomes is extremely common, and therefore it is possible for entire families to be lost from certain species or lineages just through this process of random deletion. Additionally, as transposable elements have no fixed sites, chromosomes bearing those elements may be lost from the population by selection or drift. In fact, as described in the results section, stochastic loss of transposable element families does indeed appear to be a much more frequent explanation for the observation of a patchy distribution than is horizontal transfer.

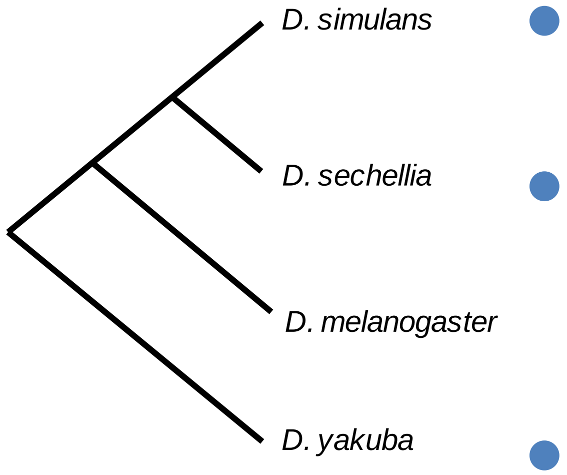


Figure 4.3: Patchy distribution of a transposable element family. Blue dots indicate species in which the family is present. Such a distribution, where a family is unexpectedly absent from a particular species, may indicate horizontal transfer, in this case, between *D. yakuba* and the ancestor of *D. simulans* and *D. sechellia*.

In broader investigations of horizontal transfer, covering a wider range of species for which divergence times are much greater, differences in codon preference and GC content can also be used to assess whether or not horizontal transfer may have occurred. In the case of the twelve *Drosophila* species investigated here, codon preference and GC content are unlikely to provide a strong enough signal to confidently infer horizontal transfer.

To be confident that horizontal transfer explains any of the observations above, all other possibilities, involving vertical transfer only, must be ruled out, including differing evolutionary rates, ancestral polymorphism, high selective constraints and stochastic losses.

The review of reported cases of horizontal transfer in *Drosophila* (Loreto et al. 2008) indicated that horizontal transfer appears to be most common for the DNA transposons. However, as this review simply collated results from reported cases, there is the possibility of bias. Negative results, i.e. investigations which suggest that no horizontal transfer has occurred, are less likely to be reported. Therefore, if there is a preference to study DNA transposons relative to the retrotransposons, or if more data on these families have been available in the past, this may result in an apparent greater proportion of horizontal transfer cases being attributed to DNA transposons. Therefore, the purpose of this study is to assess all available DNA transposon, LTR retrotransposon and non-LTR retrotransposon families listed on Repbase Update (Jurka et al. 2005), present in at least two of the twelve *Drosophila* species for which the sequenced genome is available, for horizontal transfer events. The evidence that shall be examined in this study is the amount of divergence between elements of the same transposable element family in different species, phylogenetic incongruence, and a patchy distribution across the host phylogeny. Small divergence between elements is often the most striking and convincing piece of evidence, which is not easily explained in the absence of horizontal transfer. Therefore, all families will be assessed to determine whether or not the divergence between elements in different species is ever smaller than that between the coding regions of the host alcohol dehydrogenase (*Adh*) genes. This gene has been chosen as, out of the range of possible host genes which have been used as a point of comparison in the literature when assessing putative cases of horizontal transfer (de Almeida and Carareto 2005;de Setta et al. 2007) this gene tends to have the smallest divergence between species, and therefore represents the most rigorous assessment of whether or not the transposable elements have indeed accumulated fewer mutations between them than have the host genes. This may result in possible cases of horizontal transfer being unsupported, which would have been had another host gene been selected, however, as a consequence there should be more confidence in the assumption that horizontal transfer has indeed occurred. For families in which the smallest divergence between elements for at least one interspecies comparison is found to be less than that between the *Adh* genes, phylogenetic trees will be used to try to support the case for horizontal transfer, and also to determine the most likely direction of any such transfer, that is, which was the donor and which was the recipient species. The distribution of the families across the *Drosophila* phylogeny may also go some way to support each case. Overall, this investigation will yield a conservative, unbiased indication of how frequent horizontal transfer appears to be for each of the three main groups of transposable element, DNA transposons, LTR retrotransposons and non-LTR retrotransposons. It will also give an indication of the overall frequency of horizontal transfer of transposable elements in *Drosophila*, which is difficult to determine from the collation of reported cases. Horizontal transfer has been proposed to be an essential part of the lifecycle of transposable elements in *Drosophila* (Loreto et al. 2008), and therefore it may be found that the phenomenon is more common than previously thought.

In this chapter, horizontal transfer of both class I and class II transposable element families is investigated using three lines of evidence: small divergence between elements of the same family in different species, phylogenetic incongruence and a patchy distribution across the host phylogeny. The known phylogeny of the twelve Drosophila species for which the complete sequenced genomes are available, and to which phylogenetic trees of elements are compared to detect incongruence, is presented in section 4.2.1. Four previously unidentified transposable element families, which were detected throughout the course of this investigation, are then described. Results on the horizontal transfer for the DNA transposon families are presented first out of the three major classes, followed by the non-LTR retrotransposons and finally the LTR retrotransposons. At the beginning of each of these three sections, a summary of the families for which horizontal transfer is supported is given, including which lines of evidence support the inference of horizontal transfer, the estimated number of transfer events for each family, and the species suspected to have been involved in transfer, including the direction of transfer if this can be inferred. This is followed by a family-by-family description of the evidence gathered in support of horizontal transfer, grouped according to the species in which the transposable element family is found, and the number of lines of evidence in support of horizontal transfer of that family. This begins with families restricted to the close relatives D. melanogaster, D. simulans and D. sechellia, followed by those restricted to the Sophophora, those restricted to the Drosophila subgenus, and finally those distributed throughout the *Drosophila* genus. At the end of the section on non-LTR retrotransposons, the possibility of transcriptional readthrough during retrotransposition is discussed. Finally, after presenting the evidence for horizontal transfer for each family, the geographical distribution of the host species is given, and overlap of putative donor and recipient pairs is established.

4.2 Methods

The consensus sequences of DNA transposons, non-LTR retrotransposons, and LTR retrotransposons listed in Repbase Update in June 2008 were included in this study. Searches were performed with the local alignment search tools BLAT and FlyBase BLAST (Crosby et al. 2007), using each of these consensus sequences as a query. This was done to identify in which of the twelve sequenced *Drosophila* genomes each of the transposable element families is present. These species were recorded for each family. In some cases, the family appeared to be present, but with a considerably modified consensus sequence. In these cases, the complete element was obtained by extracting upstream and downstream flanking sequences and aligning them to identify the ends of the element. The ends can be determined as the flanking DNA begins to differ and no longer aligns well between different elements. Once full length sequences were obtained, a consensus sequence could be generated. Consensus sequences were produced with no ambiguous characters, e.g. W, Y, R. Instead, the nucleotide found in the majority of elements, even where other nucleotides were common, was incorporated into the consensus. This consensus sequence was then used as a query for the repeat masking program CENSOR (Kohany et al. 2006), to determine whether it shared greater similarity with any other known transposable element sequence than with the original query. The consensus was also aligned with the original consensus, and the amount of divergence determined. Where elements aligned with the full consensus across its length, with the mutations between them relatively evenly spread across the element, this element would be considered as belonging to the same family, i.e. descended from the same ancestral element as the original query, and would not be considered a novel family. Where an element was found to share greater similarity with a different family, it was excluded from the dataset. The full-length elements were extracted and included in the file with sequences more closely resembling the original consensus. Complete sets of transposable elements of each family can be found in Supplementary Data folder 1.

Once all elements of a particular family were identified and extracted from each of the genomes in which they were present, they were aligned using ClustalW (Chenna et al. 2003). In most cases, default parameters were used for the original alignment, however, in cases where large insertions or deletions were present in many of the elements, the gap extension penalty was significantly reduced to 0.05 to account for this. In order to prevent the introduction of unreasonably large gaps, in these cases, the gap opening penalty was raised to 100. This had the problem of occasionally neglecting to include some valid short insertions and deletions. For all families, alignments were performed with both the default and the adjusted parameters, and the better alignment was chosen to be adjusted manually.

In some cases, BLAST hits were obtained corresponding to very short regions of the query sequence. Where many hits were found matching the same region, this was generally indicative of the presence of a related transposable element family, which shared high identity in this region. This could be, for example, due to constraint on protein function. In these cases, flanking DNA was extracted and input into CENSOR, to identify the transposable element family involved. In some cases, CENSOR did not yield any results. In these cases, the full-length element was found by extracting flanking DNA sequences and aligning to determine the ends. The full-length sequence was then re-entered into CENSOR to ensure it did not correspond to a known transposable element in the database. Where this was the case, the sequence was then used as a query for a general BLASTN search, to determine whether any similar sequences had been entered into GenBank. Where no entries were found, the sequence was considered to be a member of a novel family of elements. The sequence was then used as a query sequence for a BLAT search of the genome in which it had been found, and all members of the new family were extracted. These elements were then aligned to produce a consensus sequence. This consensus sequence was used as a query in FlyBase BLAST to determine which other species the family was present in. These elements were also extracted.

Alignments were opened in the program MEGA (Kumar et al. 2008), where they were used to calculate divergence in the form of p distances for all pairwise sequence comparisons. p distances were used for this investigation as comparisons are being made between the divergence between a pair of transposable elements of the same family in different species, and the Adh gene between the same species. As any correction for multiple hits would increase the divergence for both comparisons, and consequently have no effect on whether or not the divergence observed between the transposable element sequences was lower than for the host genes, p distances were considered to be a sufficient measure of divergence. However, this may represent an oversimplification, as some sequences may evolve at different rates, for example due to differences in the frequency of the four nucleotides.

This may result in the p distance between two transposable elements being less than that between the Adh genes, in a case where divergence may exceed Adh if a correction is applied. The method employed here differs from that used in other recent studies of horizontal transfer in *Drosophila*, such as the work of Bartolomé et al (2009). In that study, the distribution of synonymous divergence, using Ks as a measure, between the open reading frames of transposable elements in different species, and that between 10,150 nuclear genes in different species, are compared using a Kolmogorov-Smirnov test to determine whether or not the two distributions differ significantly.

Parameters were altered such that deletions were ignored within a pairwise comparison, but sites with gaps were not ignored in all comparisons. This allowed for the inclusion of relatively short sequences. Transposable elements in *Drosophila* are susceptible to degradation, with many experiencing large internal deletions and truncations. To maximise the amount of sequence data available for comparison, such degraded elements were included in the analysis. If only full-length elements were considered, copy numbers would be extremely small in many cases, and results would not be informative. In addition, in some cases, fragmented or partial elements were found to be present in a species in which no full-length elements were found.

A table of p distances for all pairwise comparisons (both intra-specific and inter-specific) was produced in Microsoft Excel for each transposable element family present in at least two species. Each group of pairwise comparisons (e.g. *D. melanogaster* elements compared with *D. yakuba* elements) was searched to find the pair of elements (one from each of the two species under consideration) with the smallest p distance, i.e. the least amount of divergence between them. It was noted that, frequently, the smallest p distance for an interspecific comparison corresponded to a comparison of two relatively short sequences, rather than, for example, two nearly full-length sequences. In each of these cases, the two sequences did overlap in the alignment, but by a relatively (compared to the length of the family consensus) short sequence. Cases where partial sequences do not overlap at all in the alignment are indicated by the MEGA output with a question mark ("?") and therefore do not introduce any confusion. Although p distances are measures of divergence scaled by the length of the sequence (i.e. the number of mutations per base), there is still the possibility for shorter sequences to generate lower p distances due to the fact that they are short. This can be examined assuming a Poisson distribution of mutations throughout the elements. In these cases, the pairwise comparison with the smallest p distance does not correspond to the "true" smallest p distance. The Poisson correction was performed to identify the pair of elements for each interspecific comparison which had the greatest probability of genuinely being the most closely-related pair of sequences. This effect was examined on a larger scale to determine whether elements with shorter consensus sequences generally possessed a lower smallest p distance that those with longer consensus sequences.

Where the smallest p distance between sequences from a particular pair of species was smaller than the p distance for the coding region of the *Adh* gene in those two species (Figure 4.4), it is considered likely that the identity between the two sequences is due to them sharing more recent common ancestry than the host species themselves, i.e. horizontal transfer is suspected. In these cases, the alignment of the sequences belonging to each of these families was used to produce phylogenetic trees (Supplementary

Data 2). The maximum parsimony and maximum likelihood methods were used to produce these phylogenies. In the case of the LTR retrotransposons, phylogenies of both the internal sequence and the solo LTRs were produced. For the parsimony trees, the software used came from the Phylip package (Retief 2000). The *seqboot* program was used to generate 100 bootstrap replicates for the construction of the phylogenetic tree for each family. The program *dnapars* was then used to produce a collection of equally parsimonious trees. Where possible, thorough search methods were used in an attempt to produce the best possible tree. In cases where a family contains a particularly large number of elements, especially if the consensus sequence for the family is relatively long, a more heuristic search method was used. In the most extreme cases, the program was instructed to rearrange on one best tree.

mel												
sim	0.019											
sec	0.017	0.005										
yak	0.044	0.038	0.035									
ere	0.049	0.045	0.045	0.051								
ana	0.110	0.101	0.100	0.100	0.102							
pse	0.142	0.145	0.144	0.141	0.136	0.159						
per	0.139	0.141	0.140	0.137	0.135	0.158	0.004					
wil	0.208	0.203	0.201	0.204	0.204	0.225	0.174	0.173				
vir	0.237	0.231	0.231	0.237	0.231	0.252	0.224	0.221	0.238			
moj	0.220	0.216	0.216	0.220	0.220	0.226	0.205	0.207	0.220	0.144		
gri	0.246	0.239	0.239	0.242	0.235	0.252	0.226	0.227	0.231	0.176	0.176	
	mel	sim	sec	yak	ere	ana	pse	per	wil	vir	moj	gri

Table 4.1: The divergence between the *Adh* coding regions for all interspecies comparisons.

Several of the *dnapars* parameters were modified from their default values. Due to the use of an input file from *seqboot*, which contained data corresponding to 100 bootstrap replicates, the multiple datasets parameter (M) was changed to D=100. Additionally, the input order was randomised (parameter J) to improve the accuracy of the tree. One randomisation was conducted for each bootstrap replicate, using a random number seed of 111.

The *dnapars* program output a set of most parsimonious trees for each family. The program *consense* was then used to produce a consensus phylogeny from these trees.

PHYML (Guindon et al. 2005) was used to produce phylogenies using maximum likelihood. A GTR+gamma correction was applied to account for multiple hits. Parameters were estimated from the dataset using PHYML. The phylogenies generated were compared with the maximum parsimony trees generated using *dnapars*. The trees were examined for the presence of incongruence with the host *Drosophila* tree. Phylogenetic incongruence is reported where both the parsimony and maximum likelihood trees both display a particular, well-supported incongruent relationship. In these cases, the phylogenetic incongruence was considered to be reliable, and may provide further evidence to support the hypothesis of horizontal transfer. The phylogenies produced by both methods were unrooted. A transposable element family closely related to the family under consideration, such as various members of the Gypsy superfamily of LTR retrotransposons, cannot be reliably aligned and used as an outgroup. Although sequence similarity is high enough that families can be deduced to be related, these families tend to be extremely divergent, and may only align over a short stretch of sequence. The divergence among the elements in the tree is also likely to be relatively small, making such a divergent outgroup inappropriate. The branch lengths in the tree would be extremely short compared with the length of the branch leading to the outgroup. In cases where the transposable element family is present in more than two species, theoretically, the outgroup of those host species could be used to root the tree. However, this would make the assumption of entirely vertical transmission, which cannot be assumed.

In analysing the phylogenies produced for evidence of horizontal transfer, incongruence involving extremely closely-related species, i.e. *D. persimilis* and *D. pseudoobscura*, and *D. simulans* and *D. sechellia*, was almost consistently observed for all trees containing elements from these species. This can be explained by the very close relationships between the host species and therefore has not been considered to support the hypothesis of horizontal transfer. Trees for which the only examples of incongruence involve either of these pairs of species, e.g. elements from *D. simulans* found within a clade of elements from *D. sechellia*, are considered to be congruent with the host phylogeny.

4.2.1 *Drosophila* phylogeny

The phylogeny of the twelve *Drosophila* species for which the complete sequenced genomes are available was taken from FlyBase, the source of the sequence data and the BLAST tool used to identify transposable element sequences. This phylogeny was used to determine the relationships between the twelve species, such that instances of phylogenetic incongruence could be determined when constructing phylogenies using transposable element sequences. In the past, phylogenies of the *Drosophila* genus have been produced which represent *D. yakuba* as being more closely related to *D. melanogaster* than to *D. erecta*, whereas recent phylogenies, including that presented by the *Drosophila* 12 Genomes Consortium (Clark et al. 2007), have suggested that in fact *D. yakuba* and *D. erecta* may be each other's closest relatives out of the twelve *Drosophila* species. The results are interpreted with caution where phylogenetic incongruence is observed involving these two species. However, this uncertainty will have no implications for the assessment of whether or not horizontal transfer has occurred, as this is based on the divergence between elements being less than that between the *Adh* coding regions, which is independent of the relationships between the host species.

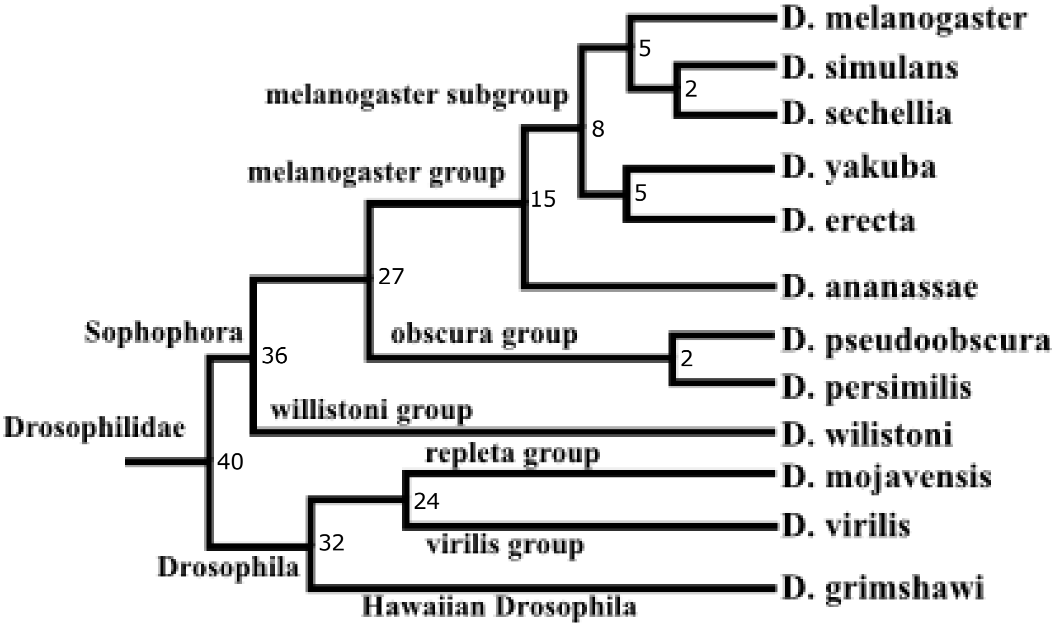


Figure 4.4: Relationships between the twelve *Drosophila* species for which the complete sequenced genome is available, adapted from FlyBase. Approximate divergence times, in millions of years, are given on each node.

4.3 Results and Discussion

4.3.1 Identification of new transposable element families

By the method described above, several previously unreported transposable element families were identified. Two novel DNA transposon families, Transib1_Dmoj/Dwil, and Helitron1_Dmoj, and two novel LTR retrotransposon families, Nobel and Gypsy_DG, were identified. This occurred where the novel elements shared regions of homology with known sequences. In the case of the DNA transposons, the homology was found in the region of the sequence corresponding to the open reading frame of the transposase protein. The homology shared by the two families in this region may therefore be due to constraint on the sequence due to transposase function. No previously unreported families of non-LTR retrotransposons were detected.

4.3.1.1 Transib1_Dmoj

The first novel family to be identified has been named Transib1_Dmoj (Styles 2008b) in accordance with the nomenclature guidelines for new transposable elements (Jolanta Walichiewicz, personal communication). This family was identified in *Drosophila mojavensis* following a BLAT search using the Transib5 consensus sequence. Eleven hits were obtained which corresponded to sequences sharing a region of homology with Transib5. Flanking regions were extracted and aligned to determine the 5’ and 3’ ends of the novel element. Multiple alignment of the eleven extracted copies was performed manually to derive a consensus sequence, using BioEdit. Of the eleven Transib1_Dmoj elements present in *D. mojavensis*, only five are full-length. Two elements are truncated at the 5’ end only, one is truncated at the 3’ end, and a further two are truncated at both ends. A single element contains a large internal deletion. Nine of the elements share very high percentage identity, between 99.84% and 100%, with the consensus sequence, with an average of 99.93%, suggesting that this family is currently active in *D. mojavensis*. The remaining two elements are more divergent, with 94.5% and 96.07% identity to the consensus. These elements may have been present in the genome for a more extended period of time, and therefore may indicate that this family is not a recent addition to the genome of *D. mojavensis*.

The consensus sequence of Transib1_Dmoj is 3058bp long. The consensus includes a transposase gene, between positions 499 and 2547. The transposase gene was identified using the ORF finder tool, followed by alignment and comparison with other Transib transposase genes to confidently identify the appropriate start codon. The Transib1_Dmoj transposase protein is 682 amino acids in length, encoded by a single 1956bp open reading frame. Transib1_Dmoj can therefore be assumed to be an autonomous family of elements, capable of mediating its own transposition. The protein shares some homology with other Transib transposases. The full-length transposase open reading frame is only found in the five full-length Transib1_Dmoj elements. The open reading frame in one of these five elements is interrupted by a 2028bp insertion, and can therefore be assumed to be non-functional, although this insertion would not introduce a frameshift into the sequence. However, as no other elements contain this insertion, it is clear that even if this element is capable of autonomous transposition, it has not yet mobilised. The open reading frame of this element, and all other full-length elements, contain numerous premature stop codons and frameshift mutations, which would render them non-functional. Therefore it appears that although these Transib1_Dmoj elements appear to have integrated into the genome recently, the elements have already lost the capacity to undergo transposition. Even the elements which share 100% identity to the consensus sequence, which does encode functional transposase, contain small deletions which render the open reading frame non-functional. Therefore, it is likely that the Transib1_Dmoj family will eventually be lost from the *D. mojavensis* genome, unless non-autonomous transposition through use of a related Transib transposase is possible.

Alignment of the Transib1_Dmoj transposase open reading frame with that of Transib5 generates 587 aligned positions, of which 220 are identical in both elements, giving a percentage identity of 37.5%. A highly conserved region of the protein contains 173 identical amino acids over 360 positions (48%). This includes a perfectly conserved string of nine amino acids (PSSTRYCRP). Whether this sequence has any functional importance is unknown. The best hit obtained from a BLASTP search using Transib1_Dmoj transposase as a query was for a transposase found in *Helicoverpa zea*, a moth more commonly known as corn earworm. This sequence shared 41% amino acid identity with Transib1_Dmoj transposase. It is possible that the homology between Transib1_Dmoj transposase and other Transib transposases may be due to functional constraint rather than relatively recent common ancestry, as suggested by the homology with *H. zea* transposase. This relatively low level of identity makes the possibility of non-autonomous transposition of Transib1_Dmoj less likely.

In *D. mojavensis*, full-length Transib1_Dmoj elements possess almost perfect 40bp terminal inverted repeats (TIRs), with only one mutation between them. TIRs have an important function in transposition. The sequence of the TIRs of Transib1_Dmoj shares homology with those of other Transib elements. Therefore this feature, along with the homology between the transposase genes, and the absence of any reasonable homology with other transposase sequences from DNA transposons in *Drosophila*, has led to the classification of this new family within the Transib superfamily.

As described above, one full-length Transib1_Dmoj element contained a transposase open reading frame interrupted by a large insertion. A BLAT search of the *D. mojavensis* genome using this insertion sequence as a query revealed many hits sharing relatively high percentage identities. As the sequence had perfectly interrupted the Transib1_Dmoj element, it was likely itself to be a transposable element. This is also supported by the presence of multiple copies of the sequence. The element was assessed for features of transposable elements. It does not possess terminal inverted repeats, does not appear to generate target site duplications, and does not encode any proteins. However, it does share a section of homology with *Drosophila* helitrons, a series of 8bp direct repeats at the 5' end. It has inserted between T and A nucleotides in Transib1_Dmoj, which is characteristic of helitrons. It is possible that this element represents a non-autonomous helitron. This element has been named Helitron1_Dmoj in accordance with standard nomenclature, and is described below.

4.3.1.2 Transib1_Dwil

The Transib1_Dmoj consensus was used as a query sequence to search the other eleven sequenced *Drosophila* species genomes, to determine whether or not the family, or any closely-related families, were present in any other species. Sequences similar to Transib1_Dmoj were found to be absent from all species except for *D. willistoni*, where multiple hits were obtained. No sequences resembling Transib1_Dmoj were identified in *D. virilis*, which is the closest relative of *D. mojavensis* for which the genome has been sequenced, suggesting either than Transib1_Dmoj was introduced into *D. mojavensis* following its divergence from *D. virilis*, around 24 million years ago, or that the family has been eliminated from the genome of *D. virilis* by stochastic loss.

27 copies of the family, referred to as Transib1_Dwil in this species (Styles 2008c), were extracted from the *D. willistoni* genome sequence. A consensus

sequence for Transib1_Dwil was obtained by multiple manual alignment of these 27 sequences in BioEdit. This yielded a 3060bp consensus sequence, which, like Transib1_Dmoj, possesses 40bp TIRs. Transib1_Dwil encodes a 672 amino acid transposase, ten amino acids shorter than that found in Transib1_Dmoj. Therefore the family is, theoretically, capable of autonomous transposition in *D. willistoni*. The transposase is encoded by a single open reading frame between positions 499 and 2518. Interestingly, two point mutations at positions 847 and 848 of the Transib1_Dwil consensus sequence, which introduce a premature stop codon, are found in six of the elements. It is unlikely, particularly given the high level of similarity between the sequences in *D. willistoni*, that these mutations have happened multiple times in parallel. Therefore, as the mutations would render the transposase non-functional, truncated to only 115 amino acids, it appears that these elements have propagated non-autonomously. Functional copies of Transib1_Dwil would be able to provide the transposase protein, which would recognise the TIRs of the mutated copies and mediate their transposition. However, since the integration of these elements, deletion events have removed sequences including the TIRs from them, such that only three possess a single intact TIR, and none have an intact pair of TIRs, therefore no further proliferation of elements containing the stop codon mutation can occur.

The maximum percentage identity between Transib1_Dmoj and Transib1_Dwil is around 95%. The identity between the consensus sequence of Transib1_Dmoj and Transib1_Dwil is 87%. This discrepancy is due to the fact that the pair of elements which share very high identity do not overlap in the 3’ end of the element. This region is relatively distinct between Transib1_Dmoj and Transib1_Dwil, compared with the majority of the element. Percentage identity drops to around 60% for a region spanning approximately 300bp.

D. willistoni belongs to the Sophophora subgenus, whereas *D. mojavensis* belongs to the Drosophila subgenus, and the two species diverged around 40mya. Therefore it is highly unlikely that Transib1_Dmoj/Dwil has been vertically transmitted, and has been retained by these two species, but lost by the other ten species for which the sequenced genomes are available. This is supported by the high percentage identity between *D. mojavensis* and *D. willistoni* elements. Therefore, it is likely that this family has been horizontally transferred between the two species, but not in the very recent past. The two species do indeed overlap in their geographical distribution, therefore this is a logical possibility. Examination of Ka and Ks values for the transposase genes between Transib1_Dmoj and Transib1_Dwil further suggest horizontal transfer. The Ka and Ks values do not differ greatly (Ka = 0.1034, Ks = 0.1709), suggesting that the similarity between elements is not a product of selective constraint. Additionally, Ks is over an order of magnitude smaller than that of *Adh* between *D. mojavensis* and *D. willistoni*, which is 1.2272. This suggests that the time to common ancestry of the Transib1_Dwil/Dmoj elements in the two species is much shorter than the time to common ancestry of the two species themselves. The possibility of horizontal transfer of Transib1_Dmoj/Dwil will be discussed further in the following section.

4.3.1.3 Helitron1_Dmoj

Helitron1_Dmoj was identified as an insertion into a Transib1_Dmoj element named “moj2” in *D. mojavensis*. It can be assumed that Helitron1_Dmoj is active, and has been transposing very recently (Styles 2008a). The divergence between all nine Transib1_Dmoj elements in *D. mojavensis* is very small, suggesting they represent very recent insertions. If this is the case, the Helitron1_Dmoj insertion into “moj2” must be even more recent. However, the highest identity of another Helitron1_Dmoj element to the insertion into “moj2” is 99.1%, which may be expected to be higher given how recent the insertion is inferred to be. However, it is possible that the source of the transposition of the “moj2” sequence is simply not represented in the genome assembly. The presence of an insertion into a recently-inserted Transib1_Dmoj element strongly suggests that Helitron1_Dmoj is currently active in *D. mojavensis*.

The “moj2” insertion was used as a query sequence for a BLAT search of the *D. mojavensis* genome sequence, which yielded 67 hits. The consensus sequence of Helitron1_Dmoj was obtained by extracting and manually aligning these 67 copies using BioEdit. The sequences directly flanking Helitron1_Dmoj sequences were extracted to reconstruct the target sites into which the elements had inserted. It was found that Helitron1_Dmoj inserts preferentially into 5’-TT-3’ target sites present in T-rich regions. Helitron1_Dmoj does not appear to produce target site duplications upon insertion, although it is possible that they have since mutated. This is unlikely to be the case, however, as the elements share a high percentage identity of, on average, 97.5% to the consensus, suggested relatively recent origin.

Helitron1_Dmoj was found to be absent from the other eleven species of *Drosophila* with sequenced genomes, including *D. virilis*, the closest relative of *D. mojavensis* for which the genome has been sequenced. This is, however, not particularly surprising, as the time to common ancestry of these two species is in excess of 20 million years. Although this family was absent from the other species for which complete sequenced genomes are available, a BLASTN search using the Helitron1_Dmoj consensus as a query reveals shorter sequences with very high identity to Helitron1_Dmoj in other members of the *repleta* group, *D. buzzatii* and its sibling species *D. koepferae*. Only one copy of the sequence is found in the available genomic data for each species, however, only a small fraction of the genomes of these species is available in GenBank. Therefore, it is possible that Helitron1_Dmoj, or a related element, is present at higher copy number in each of these species, which may be revealed should the complete genome sequence become available.

4.3.1.4 GypsyDG

An LTR retrotransposon with homology to Gypsy-like sequences was detected in *Drosophila grimshawi*, upon searching the genome with the consensus sequences of related families, which shall be referred to as GypsyDG (Styles 2009a;Styles 2009b). Six copies of GypsyDG were identified, and aligned to produce a consensus. The consensus is 3090bp in length, and contains a single open reading frame encoding a *gag-pol* polyprotein between positions 1082 and 2743. Flanking LTRs were compared for all six elements to estimate the time since integration. The maximum identity between flanking LTRs was 98.4%, suggesting relatively recent mobilisation of GypsyDG has occurred. GypsyDG does not appear to be currently mobilising in *D. grimshawi*. Of the six elements, the *gag-pol* open reading frames of four contain large deletions, and the other two are saturated with premature stop codons. It therefore appears that GypsyDG has recently lost the ability to retrotranspose.

4.3.1.5 Nobel

Nobel is an LTR retrotransposon belonging to the Bel superfamily found in *Drosophila yakuba*, *D. ananassae*, *D. pseudoobscura* and *D. persimilis*, with a copy number of 11, 54, 2 and 53 elements, respectively (Austin and Styles 2009a;Austin and Styles 2009b). The consensus sequence is 6294bp in length. Nobel encodes a single *gag-pol* polyprotein which starts at position 1107 and ends within the LTR, which is interrupted by a premature stop codon between positions 3534 and 3536 in the consensus sequence. The open reading frame is only intact in one copy of Nobel in *D. persimilis*. Flanking LTRs can be up to 100% identical suggesting that Nobel is still actively retrotransposing in *D. persimilis*. There is some substructure to this family, with a group of five elements in *D. persimilis* sharing 37 mutations from the consensus, and a group of five elements in *D. ananassae* sharing five mutations from the consensus. Despite a patchy distribution across the *Drosophila* phylogeny, there is no evidence for horizontal transfer of Nobel between the four species in which it is found.

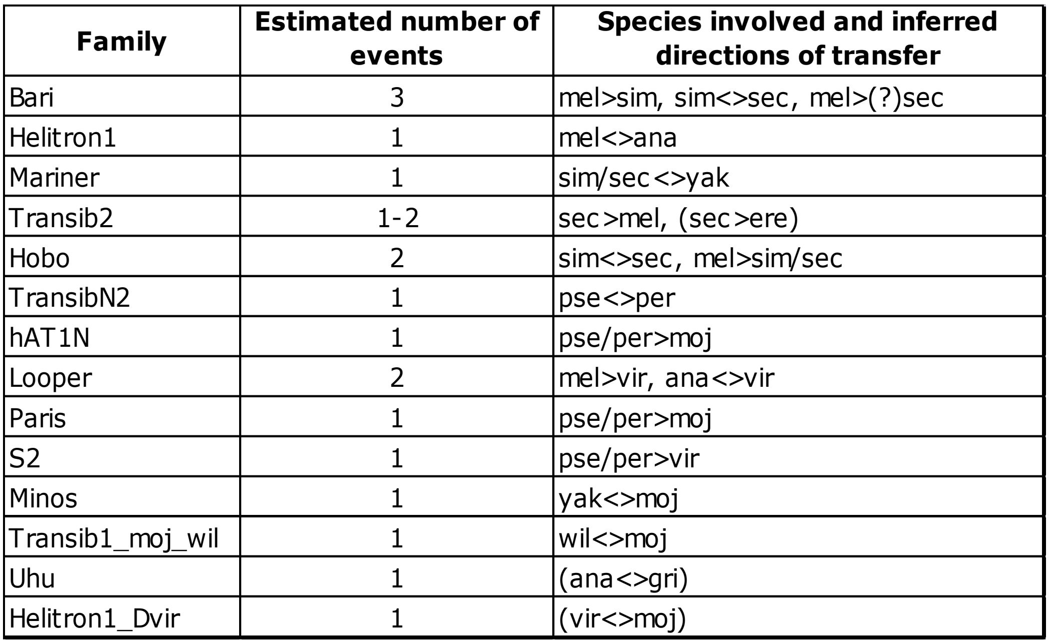
4.3.2 DNA transposons

The consensus sequences of each of the *Drosophila* transposable element families present in Repbase Update was used to determine in which of the twelve sequenced genomes each family is present. For each family, elements were then extracted from those species and aligned. The alignment was then used to calculate the smallest p distance, which represents the least divergence, or maximum percentage identity, observed between two sequences of the same family from different species (Appendix 3). 41 families of DNA transposons were found to be present in at least two of the twelve species with sequenced genomes, and were therefore investigated. Where the smallest p distance between a pair of elements, one from each species, is smaller than that observed for the coding region of *Adh*, horizontal transfer is inferred to have occurred.

In total, twelve DNA transposon families, including the newly discovered family Transib1_Dmoj/Dwil, were identified for which at least one interspecific comparison yielded divergence smaller than that between the host *Adh* genes. Six of these families are restricted to the Sophophora subgenus, and the other six are distributed across both the Sophophora and Drosophila subgenera. There are no families for which the divergence between elements in different species is smaller than for *Adh* which are found only among the species of the Drosophila subgenus. Horizontal transfer of each of these twelve families is implicated by small divergence, in addition to, in some cases, additional evidence in the form of phylogenetic incongruence, patchy distribution, or both (Table 4.2). Divergence data for the remaining DNA transposon families, for which evidence for horizontal transfer was not found, are given in Appendix 3.

Family	**Low divergence**	**Patchy distribution**	**Phylogenetic incongruence**
Bari	Yes	Yes	Yes
Helitron1	Yes	Yes	No
Mariner	Yes	Yes	No
Transib2	Yes	No	Yes
Hobo	Yes	No	No
TransibN2	Yes	No	No
hat1N	Yes	Yes	Yes
Looper	Yes	Yes	Yes
Paris	Yes	Yes	Yes
S2	Yes	Yes	Yes
Minos	Yes	Yes	No
Transib1_moj/wil	Yes	Yes	No
Uhu	No	Yes	No
Helitron1_Dvir	Yes*	No	Yes

Table 4.2: Summary of the evidence for horizontal transfer among the DNA transposons. The asterisk (*) indicates that although divergence between Helitron1_Dvir elements is small, it does slightly exceed the divergence between the *Adh* coding regions.



Family	Estimated number of events	Species involved and inferred directions of transfer
Bari	3	mel>sim, sim<>sec, mel>(?)sec
Helitron1	1	mel<>ana
Mariner	1	sim/sec<>yak
Transib2	1-2	sec>mel, (sec>ere)
Hobo	2	sim<>sec, mel>sim/sec
TransibN2	1	pse<>per
hAT1N	1	pse/per>moj
Looper	2	mel>vir, ana<>vir
Paris	1	pse/per>moj
S2	1	pse/per>vir
Minos	1	yak<>moj
Transib1_moj_wil	1	wil<>moj
Uhu	1	(ana<>gri)
Helitron1_Dvir	1	(vir<>moj)

Table 4.3: Summary of inferred horizontal transfers of DNA transposons among the twelve *Drosophila* species. The estimated number of events given is a minimum number of events required to explain the observations made. Where multiple explanations are possible, the most parsimonious, i.e., that requiring the fewest horizontal transfer events, is presented. > indicates transfer from the first species to the second. <> indicates a transfer of unknown direction. ( ) indicates lack of certainty that a transfer event has occurred. >? indicates the direction of transfer is suggested, but is uncertain. Although divergence is not smaller than for the host genes for comparisons of Uhu and Helitron1_Dvir elements, there is some evidence to suggest horizontal transfer of these families may have occurred.

4.3.2.1 DNA transposon families distributed throughout the Sophophora

The six families for which there is evidence of horizontal transfer, which are restricted to the Sophophora, are Bari, Helitron1, Hobo, Mariner, Transib2 and TransibN2. Of these, horizontal transfer of three families (Helitron1, Mariner and Transib2) is supported by two pieces of evidence, and of two families (Hobo and TransibN2) is supported by small divergence alone. There is a single family, Bari, for which horizontal transfer is supported by all three lines of evidence.

Bari is a family of mariner/Tc1-like transposons, 1728bp in length, with flanking terminal inverted repeats. A Bari insertion is believed to have been exapted to perform a regulatory function in *D. melanogaster* (Gonzalez et al. 2009). Bari elements are found in the three closely related species *D. melanogaster*, *D. simulans* and *D. sechellia*, in addition to their more distant relatives *D. erecta* and *D. ananassae*. This distribution is patchy due to the absence of Bari from *D. yakuba*, however, this could be attributed to stochastic loss of Bari from this species. Divergence between elements is smaller than that between the host *Adh* genes for three interspecies comparisons, all involving the three most closely related species, *D. melanogaster*, *D. simulans* and *D. sechellia*. Therefore, the divergence data do not support horizontal transfer as an explanation for the patchy distribution of Bari across the *Drosophila* phylogeny. However, evidence for horizontal transfer among the three closely related species is strong, with strikingly small divergence ranging from as little as 0.001 between *D. melanogaster* and *D. simulans*, up to only 0.003, and from 0.003 between *D. sechellia* and both *D. melanogaster* and *D. simulans*, up to only 0.008. These values suggest that at least three horizontal transfer events have occurred, involving all three species pairs. The divergence between elements is too small to be accounted for by transfer involving the ancestor of *D. simulans* and *D. sechellia*, which diverged only 2 million years ago.

Resolution of the topology of both the maximum parsimony and maximum likelihood trees is poor, as there is very little information available to infer precise relationships due to the sequences sharing very high identity. On the maximum parsimony tree, a single element from *D. sechellia* is found within the *D. melanogaster* clade, however, this relationship is not supported on the maximum likelihood phylogeny, in which all elements from *D. sechellia* group together. The phylogenies therefore do not reliably support horizontal transfer between these two species. The placement of the elements from *D. simulans* within the *D. melanogaster* clade, which is consistent between both phylogenetic construction methods, does suggest that horizontal transfer occurred between these two species, with *D. melanogaster* as the donor and *D. simulans* as the recipient.

Horizontal transfer events involving Helitron1, Mariner and Transib2 are supported by two of the three lines of evidence. **Helitron1** is a family of 564bp non-autonomous helitrons represented in *D. melanogaster* and *D. ananassae*. The absence of this family from species which are more closely related to *D. melanogaster* than is *D. ananassae* may suggest that the family has been horizontally transferred between these two species. This theory is supported by the divergence between elements, which can be as small as 0.058, compared with 0.110 between the *Adh* genes in these two species. Six copies of the Helitron1 family are found in *D. ananassae*, but only one copy is found in *D. melanogaster*. As there is only a single Helitron1 element in *D.*

melanogaster, the congruence or otherwise of the phylogenies produced with the host relationships would be dependent upon the position of a root. Therefore, the phylogenies do not provide any further evidence in support of horizontal transfer of Helitron1, and cannot provide any indication of the direction of any such transfer.

Mariner is an autonomous member of the Tc1/mariner superfamily of transposable elements. The consensus sequence is 1286bp in length, and encodes a functional transposase. Mariner is found in the sequenced genomes of *D. simulans*, *D. sechellia* and *D. yakuba*, and is also known to be present in *D. mauritiana* (Maruyama and Hartl 1991), a close relative of *D. simulans* and *D. sechellia*. Mariner is interestingly absent from *D. melanogaster*, despite being present in its closest relatives. The distribution of this family is therefore patchy, however, absence from *D. melanogaster* may be explained by stochastic loss from this species. Alternatively, horizontal transfer between the closely related species *D. simulans*, *D. sechellia* and *D. mauritiana*, or their common ancestor, and *D. yakuba*, most likely from *D. yakuba* into the ancestor of the other three species, would also explain the absence of Mariner from *D. melanogaster*. The Mariner family has previously been reported to have been involved in a horizontal transfer event between the *D. simulans* complex and *D. yakuba* (Lohe et al. 1995). Such a transfer is also supported by the results of this study, as divergence between elements in different species is smaller than for the *Adh* gene for the comparisons of both *D. simulans* and *D. sechellia* with *D. yakuba*, with smallest divergence of 0.012 and 0.013, respectively. Such a transfer is likely to have involved the ancestor of these two species. The phylogenies produced are congruent with known host relationships, with all elements from *D. yakuba* clustering together to the exclusion of elements from *D. simulans* and *D. sechellia*. Precise relationships are consistent between the two phylogenetic construction methods. The congruence of the phylogenies therefore does not provide evidence to suggest horizontal transfer involving *D. simulans* and *D. sechellia* since they diverged from each other. The trees cannot reveal whether or not transfer involving the ancestor of these two species occurred, as this would result in a congruent phylogeny as is observed, which cannot be distinguished between that which would be observed had no horizontal transfer occurred. The small divergence between elements does however suggest horizontal transfer has occurred, and the trees can be taken to support such a transfer involving the ancestor of *D. simulans* and *D. sechellia* rather than occurring in the very recent past. Mariner is also reported to have undergone many more horizontal transfer events (reviewed in Loreto et al. 2008), involving species for which the complete genome sequences are not available, for example between the *montium* subgroup of *Drosophila* and *D. vallismaia*.

Transib2 is an autonomous member of the Transib superfamily, and is represented by a 2844bp consensus sequence. Transib2 is found in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*. There is therefore no evidence from the distribution of these elements to suggest horizontal transfer. However, the smallest divergence in comparison of Transib2 elements from *D. melanogaster* and *D. sechellia* is only 0.010, which is smaller than for the *Adh* genes between these species, suggesting that horizontal transfer may have occurred between them. The phylogenies produced are consistently incongruent with the host tree, although some precise relationships are inconsistent between the two methods. Both trees support a horizontal transfer event involving *D. melanogaster* and *D. sechellia*, as on both trees, elements from *D. melanogaster* are found scattered throughout the *D. sechellia* clade, forming closer relationships with the elements from this species than do those from its close relative *D. simulans*. These relationships are well-supported by both phylogenetic reconstruction methods. This suggests a recent transfer event has occurred from *D. sechellia* into *D. melanogaster*. The phylogenies also suggest horizontal transfer may have occurred from *D. sechellia* again as the donor species, to *D. erecta* as the recipient. On the maximum likelihood tree, some of the *D. erecta* elements are scattered throughout the *D. sechellia* clade. On the maximum parsimony tree, these elements from *D. erecta* group together, but this group is found within the *D. sechellia* clade. Such a transfer cannot be refuted by the divergence data, as the smallest divergence between Transib2 elements in *D. sechellia* and *D. erecta* is 0.052, slightly larger than the corresponding divergence between the *Adh* coding regions, which is 0.045.

Horizontal transfer of Hobo and TransibN2 is supported by only a single piece of evidence: that the divergence between elements in different species can be smaller than that between the host *Adh* genes for the same interspecies comparison.

Hobo is a 3016bp long autonomous member of the hAT superfamily. It is found in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*. There is therefore nothing concerning its distribution across the *Drosophila* phylogeny to suggest the Hobo family has been involved in horizontal transfer. Hobo is an interesting family to investigate in terms of its evolution, as like Nomad of the LTR retrotransposons, which is discussed in detail below, many elements appear to have propagated non-autonomously. There is a mixture of full-length elements and copies consisting only of the extreme 5’ and 3’ ends, with the middle portion of the element absent, in all species with the exception of *D. erecta*. These elements have clearly retained the sequence information required to be recognised by the transposase protein encoded by full-length Hobo elements, which has led to their continued proliferation (Depra et al. 2009). It can be assumed that an element of this type would not be able to undergo horizontal transfer into a recipient species from which full-length Hobo elements were absent. Hobo has been reported to have been introduced into the ancestor of *D. melanogaster* and *D. simulans* by horizontal transfer (Depra et al. 2009), however, the extent of the divergence between Hobo elements in *D. melanogaster*, *D. simulans* and *D. sechellia* suggests that Hobo has also been horizontally transferred among these species. The smallest divergence observed for all three interspecies comparisons is extremely small, ranging from 0.000, whereby elements are identical to each other in *D. simulans* and *D. sechellia*, to 0.001 for comparisons between both of these species and *D. melanogaster*. These values are strikingly small and strongly support the hypothesis of recent common ancestry of these elements, and therefore horizontal transfer. Variation among the elements in the different species, however, is quite high, suggesting any horizontal transfer event did not introduce Hobo into these species. This is also supported by the more extensive distribution of Hobo, which is also found in *D. yakuba* and *D. erecta*, and can be assumed to have been inherited by the other three species by vertical transmission from the common ancestor. The divergence between elements in different species is so small that it cannot be attributed to transfer events involving the ancestor of *D. simulans* and *D. sechellia*. At least three individual transfer events may therefore have occurred.

The phylogenetic tree of Hobo elements produced using the maximum likelihood method is incongruent with known host relationships, however, this is inconsistent with the relationships present in the maximum parsimony tree, which is congruent with the host phylogeny. Therefore, this cannot be taken as convincing evidence for horizontal transfer. However, on the maximum likelihood tree, all elements from *D. melanogaster* form a clade together within the clade of elements from *D. simulans* and *D. sechellia*. This would suggest transfer from either *D. simulans* or *D. sechellia* into *D. melanogaster*. A second transfer between *D. simulans* and *D. sechellia* would account for the high degree of similarity between all three species, and therefore only two horizontal transfer events are required to explain the divergence data, if the relationships supported in this phylogeny are genuine.

TransibN2 belongs to the Transib superfamily, and is found in *D. pseudoobscura* and *D. persimilis*. The smallest divergence observed by comparing TransibN2 elements from *D. pseudoobscura* with elements from *D. persimilis* is 0.000, that is, 100% identity between the elements, compared with divergence of 0.005 for the *Adh* coding sequence. The TransibN2 consensus is only 40bp in length, and therefore divergence between elements smaller than between the *Adh* genes is not particularly surprising. The two species are very closely related, and it is possible that there are no mutations between some of these sequences simply by chance. *D. pseudoobscura* and *D. persimilis* diverged around 2 million years ago, and all DNA transposons included in this study which were present in one of those two species were also found in the other. It is very likely that these two species possess TransibN2 elements in their genomes due to vertical transmission. However, it is possible that horizontal transfer may have occurred, although it is unlikely to have introduced the family. The phylogenies of TransibN2 elements produced are incongruent with host relationships, in that elements from the two species do not form distinct clades on the trees. Close relationships between elements from the two species are well-supported on both trees. Again, this is expected due to the close relationship between the host species, and therefore does not provide convincing evidence for horizontal transfer. Therefore, it appears that the similarity between elements of this family in these two species can be attributed to the short divergence time of the host species and the extremely short length of the elements, rather than horizontal transfer. It is possible that horizontal transfer has occurred, but it would be impossible to identify convincing evidence, and therefore such an event would remain undetected. However, hybridisation between *D. pseudoobscura* and *D. persimilis* within the last 200,000 years (Kulathinal et al. 2009) makes horizontal transfer of transposable elements through introgression a realistic possibility, and therefore identical TransibN2 elements may be present in these species as a result of this process.

4.3.2.2 DNA transposons distributed throughout the Sophophora and Drosophila subgenera

In the case of six out of the eleven DNA transposon families distributed throughout the Sophophora and Drosophila subgenera, for at least one interspecies comparison, the smallest divergence observed between elements is less than the corresponding value for the coding region of *Adh*. These are hAT1N, Looper, Minos, Paris, S2 and Transib1_Dmoj/wil. The distribution of smallest p distances between the *Drosophila* and Sophophora subgenera are shown in Figure 4.5.

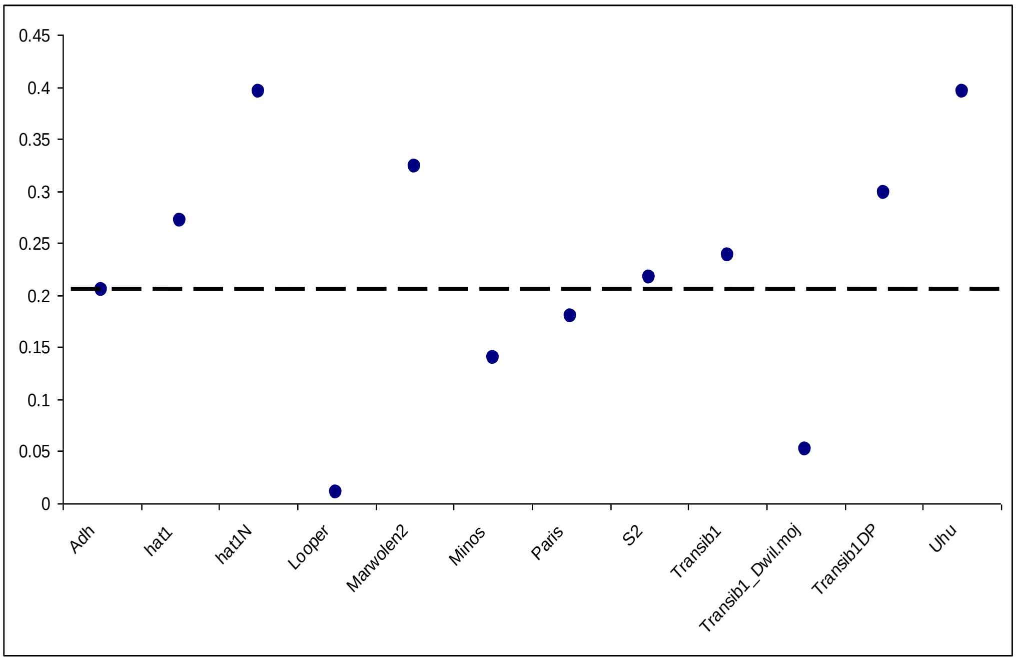


Figure 4.5: Graph showing the smallest divergence observed in comparison of elements of a particular DNA transposon family in a species belonging to the Sophophora subgenus with elements of the same family in a species of the Drosophila subgenus. The average divergence between the *Adh* coding regions between the Sophophora and Drosophila species is also shown, and indicated as a point of comparison by the dotted line. Six families have diverged, in at least one case, to a lesser extent between the Sophophora and Drosophila than have the equivalent *Adh* genes.

Observing divergence smaller than *Adh* for such a divergent pair of species as those of the Sophophora and Drosophila subgenera, which diverged around 40 million years ago, provides more convincing evidence of horizontal transfer than for closely-related species, such as *D. pseudoobscura* and *D. persimilis*, whose elements tend to be mixed together in phylogenies. Over longer periods of time, elements are less likely to possess small numbers of mutations by chance. In these cases, particularly as the elements can be assumed to be unconstrained relative to the host *Adh* coding region, it would certainly be expected that elements following the route of vertical transmission would have accumulated more mutations per base in the same period of time than would the equivalent *Adh* gene. It may therefore be argued that the cases of putative transfer between the Sophophora and Drosophila presented below represent the most convincing cases of horizontal transfer of DNA transposons. It is also possible that, as transposable elements are eliminated so rapidly from *Drosophila* genomes, that the mere presence of a DNA transposon family in such diverged species may be indicative of horizontal transfer, particularly where the distribution of the family is patchy across the host phylogeny.

Six DNA transposon families present in both the Sophophora and Drosophila subgenera yielded a smallest divergence, for at least one interspecies comparison, smaller than the equivalent divergence between *Adh* genes. Horizontal transfer of four of these families, hat1N, Looper, Paris and S2, is supported by all three lines of evidence investigated: small divergence, phylogenetic incongruence and patchy distribution.

hAT1N is a member of the hAT superfamily, and is 557bp long. hAT1N is a short, non-autonomous version of hAT1, which is likely mobilised by the hAT1 transposase. hAT1N elements are found in *D. ananassae*, *D. pseudoobscura*, *D. persimilis* and *D. mojavensis*. hAT1 elements are found in all of these species except *D. ananassae*. It may be that hAT1 has been lost from *D. ananassae*, and consequently hAT1N may no longer be able to proliferate in this species, and therefore is also likely to be lost as elements are deleted or lost by drift over time. The distribution of hAT1N is unexpected, as it is found in three species from the Sophophora subgenus, along with *D. mojavensis* from the Drosophila. This patchy distribution of elements might be indicative of horizontal transfer from a species of the Sophophora to *D. mojavensis.* The family also appears to have been lost stochastically from the lineage leading to *D. melanogaster*, *D. erecta* and their closest relatives. The smallest divergence between elements in both species of the *obscura* group, *D. pseudoobscura* and *D. persimilis*, compared with *D. mojavensis*, was smaller than that for the *Adh* gene, at only 0.145 for the comparison with *D. pseudoobscura*, and 0.087 for the comparison with *D. persimilis*. These values, although not strikingly small, do suggest more recent common ancestry between the elements than the host species themselves, and therefore that horizontal transfer has occurred. Given the close relationship between *D. persimilis* and *D. pseudoobscura*, which only diverged around two million years ago, a single transfer can be assumed to have occurred, involving the ancestor of these two species. The phylogenies produced support transfer between these species, as the two elements from *D. mojavensis* are found within the clade of elements from *D. persimilis* and *D. pseudoobscura* (Figure 4.6). This relationship suggests that transfer occurred involving the ancestor of *D. persimilis* and *D. pseudoobscura* as the donor, and *D. mojavensis* as the recipient species, as supported by the distribution of the hAT1N family across the *Drosophila* phylogeny. All elements from *D. ananassae* form a well-supported clade in both phylogenies.

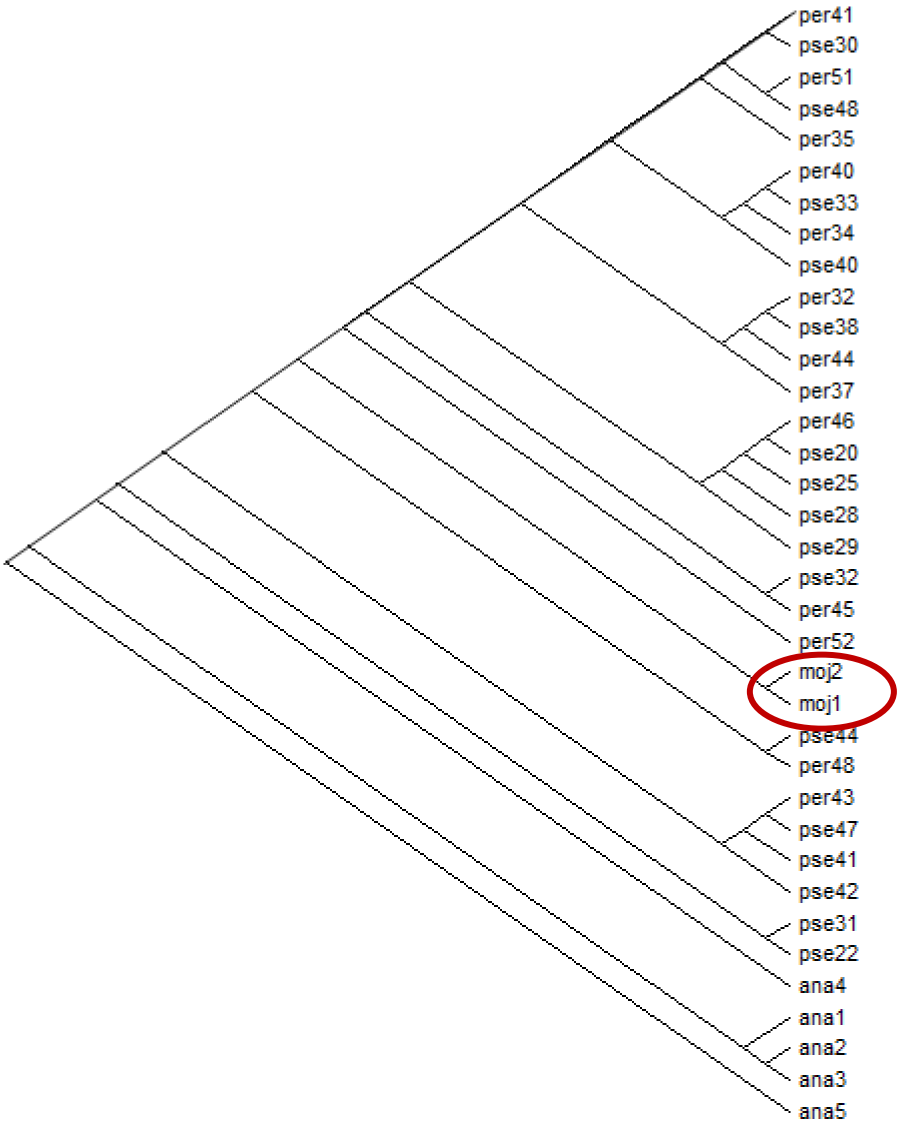


Figure 4.6: A section of the maximum parsimony phylogeny of hAT1N elements, showing the incongruent positioning of elements from *D. mojavensis*.

Looper is a family of piggyBac-like DNA transposons, and is 1881bp in length. Looper is an interesting case to investigate with respect to horizontal transfer, as it is believed that Looper may have been introduced into the *Drosophila* lineage by horizontal transfer across phyla, due to its greater similarity with piggyBac-like elements from mammals than those from other insects such as the moth (Kapitonov and Jurka 2002b). Looper elements show a patchy distribution across the *Drosophila* phylogeny, present in the closely-related species *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, and *D. ananassae*, and also the more distantly related species *D. virilis*. Interestingly, Looper is present in all members of the *melanogaster* group except for *D. erecta*. Presumably this is due to stochastic loss from the *erecta* lineage. However, the presence of Looper in *D. virilis* may be indicative of horizontal transfer, either between this species and the ancestor of the Sophophoran species in which Looper is present, or between *D. virilis* and one of the contemporary species. Divergence between elements is incredibly small between *D. melanogaster* and *D. virilis*, at only 0.01, compared with 0.237 between the *Adh* coding regions of these species. This strongly suggests that recent horizontal transfer may have occurred between these species. Divergence between elements of *D. virilis* and those of the other species is also smaller than for the equivalent host genes, however, this appears to be as a consequence of the transfer involving *D. melanogaster* and *D. virilis*. A gradual increase in the smallest divergence between each species compared with *D. virilis* is seen as the divergence from *D. melanogaster* increases (Figure 4.7). This suggests the direction of transfer is likely to have been from *D. melanogaster* into *D. virilis*. The similarity of elements from other species to those in *D. virilis* is therefore presumably a result of their close relationship with *D. melanogaster*, and not the result of further transfer, and would be consistent with vertical transmission were the elements found in *D. melanogaster* rather than *D. virilis*.

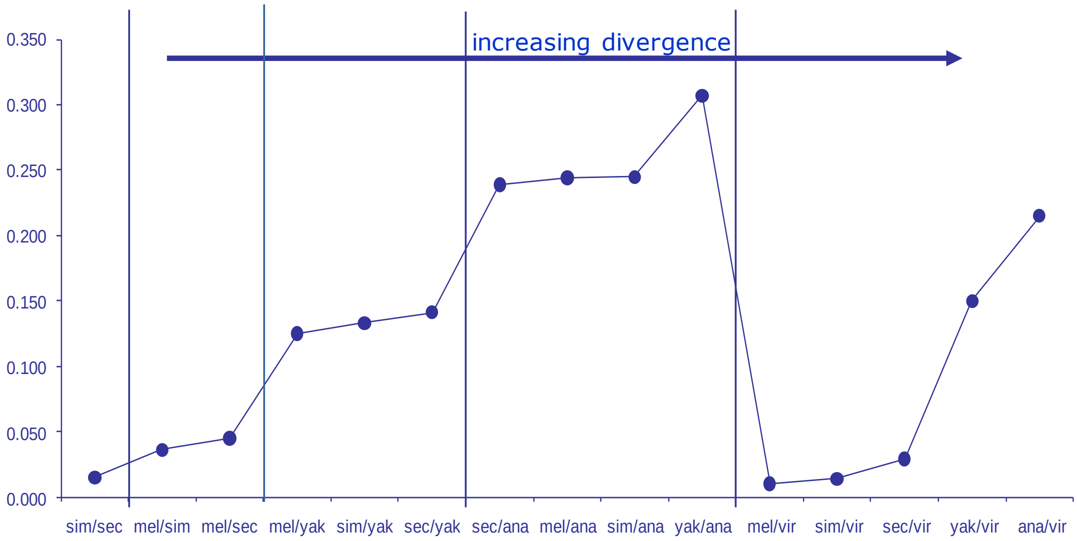


Figure 4.7: The smallest divergence observed between Looper elements for all interspecific comparisons, in order of increasing divergence time.

In support of this hypothesis, the phylogenies produced are indeed incongruent with known host relationships (Figure 4.8). An element from *D. virilis* clusters with an element from *D. melanogaster* within the *D. melanogaster* clade. This supports the hypothesis described above of transfer involving *D. virilis* as the recipient species. Another striking example of phylogenetic incongruence in the Looper phylogeny is a highly-supported clustering of an element from *D. virilis* with one from *D. ananassae*. This may represent a second instance of horizontal transfer. The divergence between these elements is 0.215, smaller than the divergence between the coding regions of the *D. virilis* and *D. ananassae Adh* genes. Therefore, it appears that a second horizontal transfer event has indeed occurred, and that the small divergence observed between *D. ananassae* and *D. virilis* is not simply an effect of the transfer from *D. melanogaster* to *D. virilis*. The direction of this transfer cannot be inferred from the phylogenies. Further incongruence which is consistent between both phylogenetic construction methods is observed which is not supported by divergence data, as a few elements from *D.*

simulans and *D. sechellia* fall within the *D. melanogaster* clade. This may suggest transfer occurred from *D. melanogaster* into the ancestor of these species.

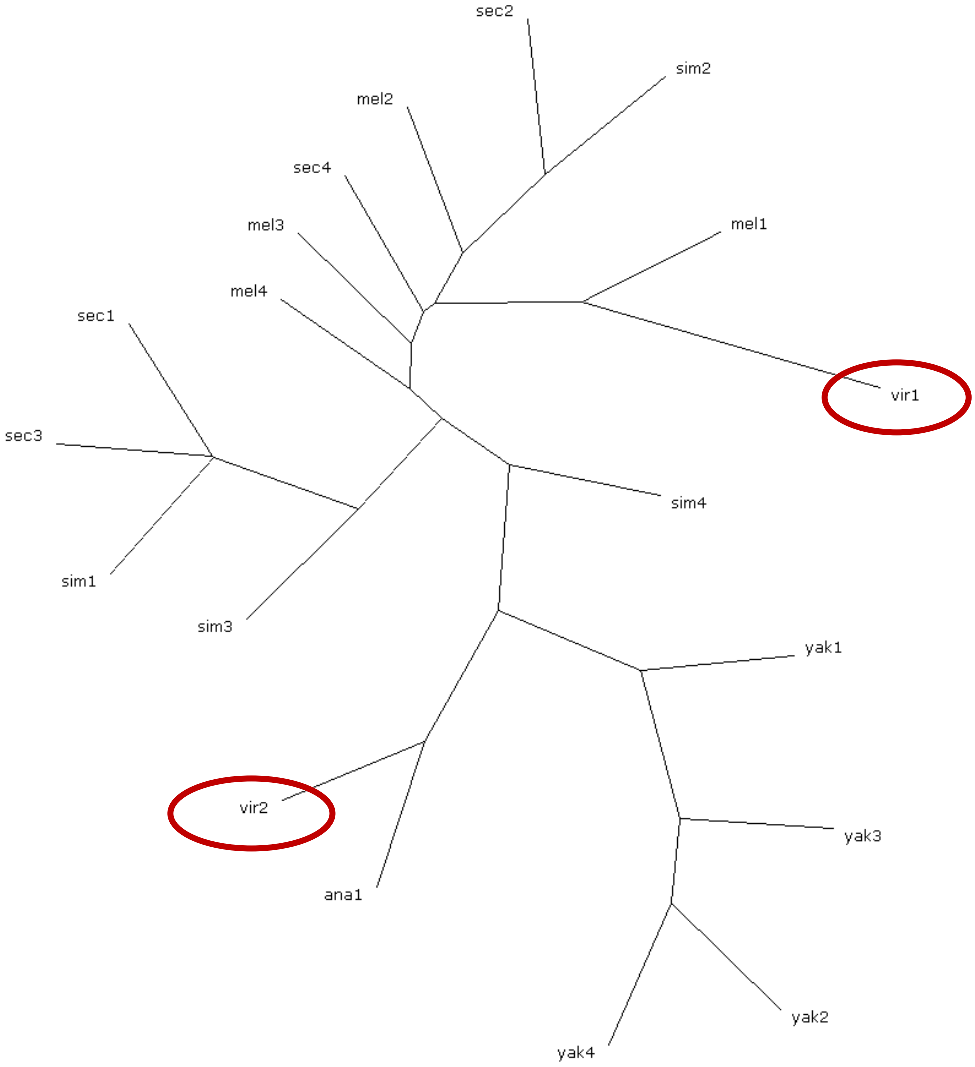


Figure 4.8: Maximum parsimony phylogeny of Looper elements, showing phylogenetic incongruence.

Paris is another member of the Tc1/mariner superfamily. It is an autonomous family, which encodes a Tc1-like transposase between positions 394 and 1440 of its 1730bp consensus. Paris has the same unusual distribution as hAT1N, as it is present in *D. ananassae*, *D. persimilis*, *D. pseudoobscura* and *D. mojavensis*, but in addition is present in *D. virilis*. This patchy distribution may be attributed to horizontal transfer between the Sophophora and Drosophila subgenera. It also appears that the Paris family, as in the case of hAT1N, may have been stochastically lost from the ancestor of *D. erecta*, *D. melanogaster* and their closest relatives. Divergence between Paris elements in different species does indeed suggest that horizontal transfer between the two subgenera may have occurred. However, divergence is only smaller than between the host *Adh* coding regions in one case, between *D. persimilis* and *D. mojavensis*. The smallest amount of divergence between Paris elements in these two species is 0.180, compared with 0.207 for the coding region of *Adh*. *D. persimilis* and *D. pseudoobscura* are very closely related to each other, having only diverged around two million years ago. It is therefore surprising that the same degree of identity is not shared between *D. mojavensis* and *D. pseudoobscura* as for *D. mojavensis* and *D. persimilis*. A recent transfer from *D. mojavensis* into *D. persimilis* would account for this, but given the divergence between elements, the transfer does not appear to have occurred recently. This case is therefore similar to hAT1N, and may be a result of the smaller copy number of elements in *D. pseudoobscura*. The phylogenies produced support the hypothesis of horizontal transfer, as the elements from *D. mojavensis* consistently cluster together within the *D. pseudoobscura*/*D. persimilis* clade, suggesting *D. mojavensis* was the recipient species (Figure 4.9). All elements from *D. virilis* form a clade together on the tree, however, unexpectedly, a single element from *D. ananassae* is found within this clade. This relationship is well-supported in both phylogenies, in 100% of bootstrap replicates on the maximum parsimony tree, and with a score of 0.987 on the maximum likelihood tree. The smallest divergence between Paris elements in *D. ananassae* and *D. virilis* is 0.273, which is not greatly in excess of the divergence between the *Adh* coding regions in these two species, which is 0.252. It is therefore possible, given the evidence from divergence and phylogenetic incongruence, that horizontal transfer of Paris has occurred twice between the Sophophora and Drosophila, with one event involving the *obscura* and *repleta* groups, and the other involving the *melanogaster* and *virilis* groups.

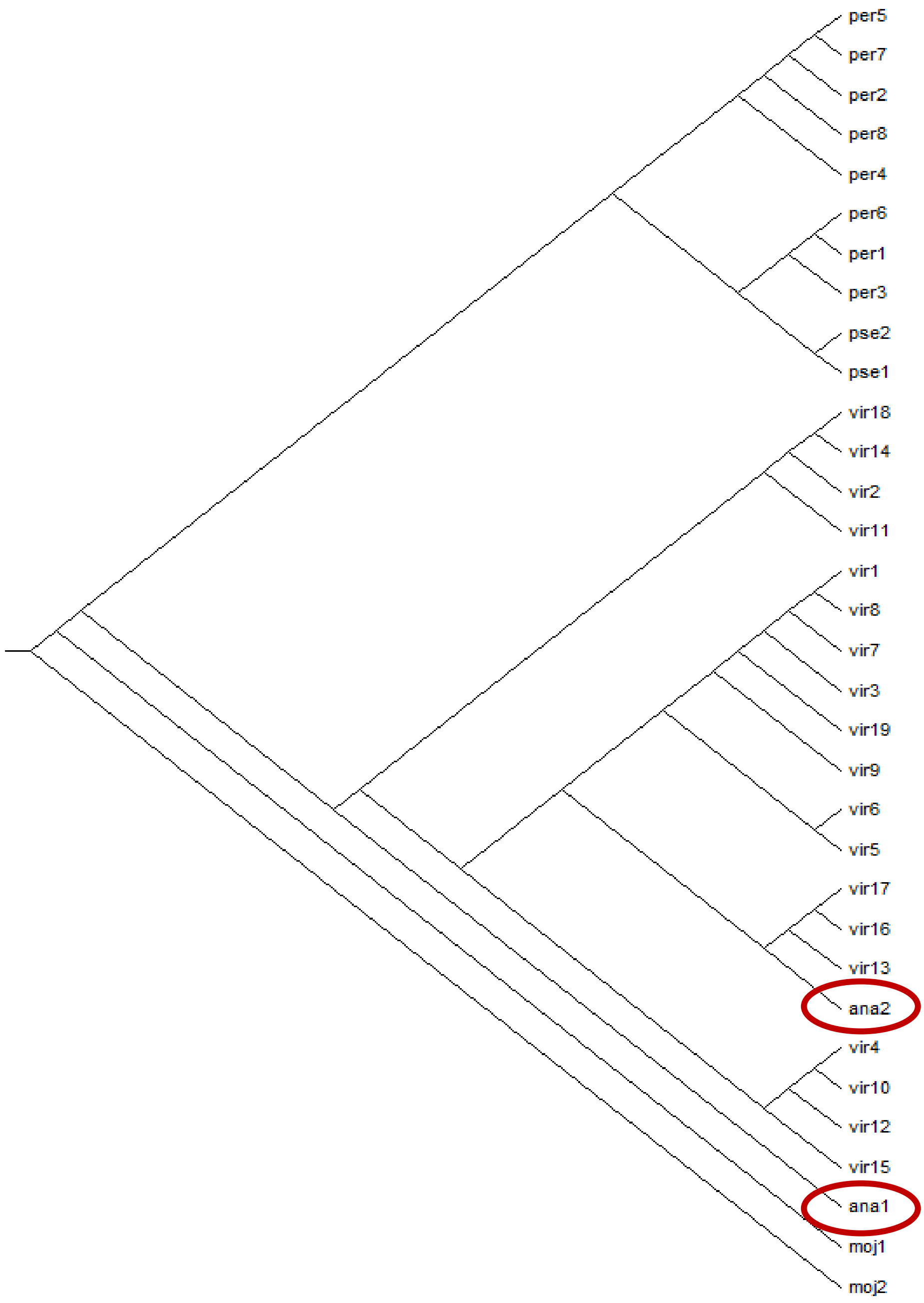


Figure 4.9: Maximum parsimony phylogeny of Paris elements demonstrating phylogenetic incongruence.

S2 is a member of the Tc1/mariner superfamily, with a 1735bp consensus sequence. S2 does not encode transposase, and therefore propagates non-autonomously. S2 is present in the complete genome sequences of three species from the Sophophora subgenus (*D. melanogaster*, *D. pseudoobscura* and *D. persimilis*) and two species from the Drosophila subgenus (*D. virilis* and *D. mojavensis*). The patchy distribution in the Sophophora, and distribution in both the Sophophora and Drosophila, could be indicative of horizontal transfer. Alternatively, S2 may be absent from the other members of the Sophophora, in particular the other members of the *melanogaster* group (*D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta* and *D. ananassae*) due to stochastic loss of the family during their evolutionary history. This would imply that S2 was present in the ancestor of the *Drosophila* genus, but has since been lost on most lineages, including, in the case of *D. simulans* and *D. sechellia*, during the last five million years. Divergence between elements supports a possible case of horizontal transfer between the Sophophora and Drosophila, as the smallest divergence observed between S2 elements in *D. persimilis* and *D. virilis* is only 0.217, smaller than the divergence between the *Adh* genes in these species. However, this level of divergence is not particularly small, and therefore does not suggest that horizontal transfer has occurred recently. This would imply that it was the ancestor of *D. pseudoobscura* and *D. persimilis*, rather than *D. persimilis* itself, which was involved in the transfer event, which is not suggested from the divergence between elements in these two species. The smallest divergence between S2 elements in *D. pseudoobscura* and *D. virilis* is 0.261, which is larger than the p distance for *Adh*, 0.224. The difference between *D. pseudoobscura* and *D. persimilis* is likely due to chance, or may be attributed to the lower copy number of S2 in *D. pseudoobscura*. However, it does raise the issue of cases of horizontal transfer perhaps not being detected, as transfers occurring further into the past are less likely to result in divergence smaller than that between the host genes. As S2 is a non-autonomous family, it can be assumed that horizontal transfer would only be followed by successful proliferation, and therefore detection, if the autonomous family responsible for its mobilisation is present in the recipient species. Paris is the only Tc1/mariner-like DNA transposon family investigated in this study which is present in both *D. persimilis* and *D. virilis*, and may therefore be capable of mobilising the S2 family. The S family is present in neither of these two species. The phylogenies produced using both phylogenetic construction methods support horizontal transfer of S2. All of the elements from *D. virilis* form a clade together, as do the elements from *D. mojavensis*. However, the elements from these two species do not group together on the trees, which is incongruent with the known host relationships, as these two species belong to the Drosophila subgenus. The elements from *D. virilis* fall within a mixed clade of elements from *D. pseudoobscura* and *D. persimilis* (Figure 4.10), supporting transfer involving the ancestor of these two species as the donor and *D. virilis* as the recipient. Further transfer is suggested by the consistent placement of two elements from *D. melanogaster* within the *D. pseudoobscura*/*D. persimilis* clade, however, such a transfer is not supported by divergence data.

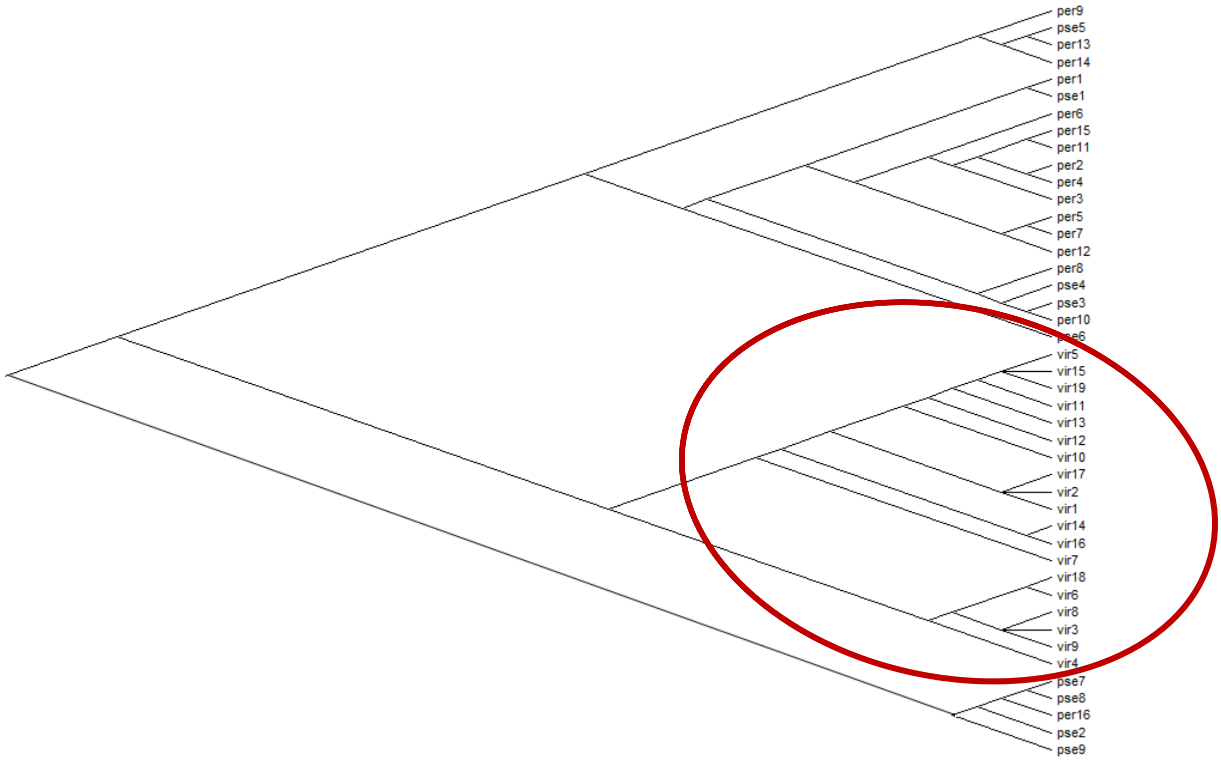


Figure 4.10: Section of the maximum parsimony phylogeny of S2 elements, showing the incongruent position of elements from *D. virilis* within the clade of elements from *D. pseudoobscura* and *D. persimilis*.

The remaining two DNA transposon families distributed across the Sophophora and Drosophila subgenera are implicated to have been involved in horizontal transfer by two of the three lines of evidence. There are no families distributed across the two subgenera for which horizontal transfer is only supported by a single piece of evidence.

Minos is a member of the Tc1/mariner superfamily. Its consensus sequence is 1775bp in length, and encodes transposase. Minos is found in the sequenced genomes of *Drosophila mojavensis* and *Drosophila yakuba*. Minos is also known to be present in *D. hydei* (Franz and Savakis 1991). *D. hydei* is a member of the *repleta* group of the Drosophila subgenus, along with *D. mojavensis*, whereas *D. yakuba* belongs to the Sophophora. Absence of

Minos from species closely related to *D. yakuba*, such as *D. melanogaster*, might suggest horizontal transfer of Minos into *D. yakuba*. Such a transfer is supported by the extent of the divergence between Minos elements in *D. yakuba* and *D. mojavensis*, which can be as little as 0.140, compared with 0.220 between the *Adh* coding regions in these species. The phylogenies produced are consistently congruent with host relationships, in that all elements from *D. yakuba* cluster together to the exclusion of all elements from *D. mojavensis*. However, this is as would be expected had transfer introduced the Minos family into *D. yakuba*. It is not possible to support this with only elements from two species available to construct the phylogenies. Such a transfer should result in elements from *D. yakuba* falling within the *D. mojavensis* cluster, which might be observed were it possible to root the tree. Horizontal transfer of Minos is reported to have occurred within the *repleta* group, and between the *repleta* group and *D. saltans* (de Almeida and Carareto 2005).

As described previously, **Transib1_Dmoj/Dwil** is implicated in horizontal transfer. The family is present in a single species of the Sophophora and a single species of the Drosophila subgenus, and therefore follows a patchy distribution across the phylogeny. The family is also absent from the closest relative of *D. mojavensis* for which the sequenced genome is available, *D. virilis*. Horizontal transfer of Transib1_Dmoj/Dwil between these two species is supported by the small divergence between elements, which can be as little as 0.052 between elements which do not overlap in the divergent 3' region described previously, or 0.136 between elements which overlap across the entire length. This is much smaller than the divergence between the host *Adh* genes, which is 0.220 between these species. The horizontal transfer event could have occurred in either direction, but is perhaps more likely to have occurred from *D. willistoni* into *D. mojavensis*. This is supported by the fact that the intra-species divergence of Transib1_Dwil elements in *D. willistoni* is much higher, suggesting the elements are older. In *D. mojavensis*, most elements share very high percentage identity (>99%). However, two divergent elements are also found, which might suggest an older “subfamily” was present and has since been replaced. This might suggest a recent invasion of this species by Transib1_Dmoj, followed by rapid proliferation in a naïve genome. However, the identity between the two species is not sufficiently high to support this theory. The higher copy number of elements in *D. willistoni* might potentially suggest the family has been there longer, however, as the intraspecific divergence is lower than the interspecific divergence, this is not really convincing evidence, therefore the direction of transfer remains inconclusive. Due to the poorly aligning region near the 3’ end, and numerous other species-specific mutations throughout the length of the element, the phylogenies produced of Transib1_Dmoj/Dwil elements are congruent with the host phylogeny, with all *D. mojavensis* elements clustered together to the exclusion of all *D. willistoni* elements. The branch separating the clade containing all *D. willistoni* elements from the clade containing all *D. mojavensis* elements is found in 100% of bootstrap replicates in the maximum parsimony tree, and has a score of 1.0 on the maximum likelihood tree.

4.3.2.3 Divergence greater than between *Adh* coding regions

Observing divergence between elements of the same DNA transposon family in different species larger than that between *Adh* coding regions cannot be used to automatically rule out the possibility of horizontal transfer, even between the species for which the genome sequence is available. As has been shown, for example in the case of Paris and hAT1N, horizontal transfer can be confidently inferred where the smallest divergence between a particular pair of species, which would be expected to be smaller than between the host genes, is actually greater. In both of these cases, this was attributed to the small copy number of elements in *D. pseudoobscura*, which eliminates transposable elements from its genome more rapidly than do the majority of *Drosophila* species for which the sequenced genome is available, including its very close relative *D. persimilis*. Investigation of some DNA transposon families suggests that some, for which the smallest divergence for all interspecies comparisons was in excess of the divergence between *Adh* coding regions, may have undergone horizontal transfer. In some cases, as in the case of Transib2 described above, this may be due to the smallest divergence only slightly exceeding the required threshold, or due to the presence of a transposable element family in distantly related species, which is absent from much more closely-related species. An example of the latter case is the family Uhu.

Uhu is a family of DNA transposons identified in the Hawaiian *Drosophila* species *D. heteroneura* (Brezinsky et al. 1990). It is also found in the genome of the Hawaiian *Drosophila* species for which the complete sequenced genome is available, *D. grimshawi*. Uhu was not found in any of the other available sequenced genomes, but a similar element was identified in *Drosophila ananassae*. The consensus sequences for the elements in both species are exactly the same length, 1655bp. Either this family has been horizontally transferred between *D. ananassae* and the Hawaiian *Drosophila* at some point in the past, or the family was present in the ancestor of the *Drosophila* genus and has been stochastically lost at least four times independently. Losses would be required from the *D. virilis*/*D. mojavensis* lineage, *D. willistoni*, the *D. pseudoobscura*/*D. persimilis* lineage, and from the ancestor of *D. melanogaster*, *D. erecta* and their closest relatives. This is a reasonable proposition; however, it is tempting to speculate horizontal transfer as *D. ananassae* does indeed share geographical overlap with the Hawaiian *Drosophila*, and transfer between these species has also been proposed for a family of LTR retrotransposons, Osvaldo, discussed below. The divergence between the *Adh* coding regions of *D. ananassae* and *D. grimshawi* is 0.252, whereas the smallest divergence between Uhu elements in these two species is 0.395, which is considerably higher. Assuming *Adh* is under greater constraint than the Uhu element, it is possible that Uhu was horizontally transferred in the distant past, and has since mutated considerably. However, such an event can be assumed to have pre-dated the speciation of *D. ananassae* and its closest relatives for which the sequenced genomes are available, such as *D. melanogaster*. Such a transfer would therefore not explain the absence of Uhu from these species. This is an ambiguous case where the distribution of elements cannot be convincingly attributed to either horizontal transfer of stochastic loss.

4.3.2.4 Divergence slightly in excess of the *Adh* coding region

There is a DNA transposon family, Helitron1_Dvir, for which smallest divergence between elements only slightly exceeds that between the *Adh* coding regions for at least one interspecies comparison, and was considerably smaller than the average smallest divergence for DNA transposons for each particular interspecies comparison. Two families for which this is also the case, Paris and Transib2, have been discussed already, as interspecies comparisons of these two families also yielded divergence smaller than between the *Adh* genes. In these two cases particularly, it might be likely that the similarity of the divergence values obtained to the value for

Adh might be due to horizontal transfer, as, due to a presumed lack of constraint on these elements, the *Adh* value should be considerably smaller, as its coding region is under constraint along its entire length.

Helitron1_Dvir belongs to the Helitron group of transposable elements. Its consensus is 8816bp long and contains a single open reading frame rendering Helitron1_Dvir capable of autonomous proliferation. This family of helitrons is only found in *D. virilis* and its closest relative for which the complete genome sequence is available, *D. mojavensis*. It is unknown whether this family is present in other members of the *repleta* and *virilis* groups. There is therefore no evidence from the distribution of Helitron1_Dvir across the *Drosophila* phylogeny to suggest it has undergone a horizontal transfer event. The smallest divergence between Helitron1_Dvir elements in *D. virilis* and *D. mojavensis* is 0.147, only slightly in excess of the value for the *Adh* coding region, which is 0.144. This therefore does not fulfil the stringent criteria to confidently conclude that horizontal transfer has occurred, but is considerably smaller than the divergence between elements of the other two DNA transposon families which are present in these two species, Paris (0.308) and S2 (0.276).

The phylogenies produced are consistently incongruent with the host phylogeny (Figure 4.11), as *D. virilis* elements and *D. mojavensis* elements do not form two distinct clades. This is a rare example of phylogenetic incongruence being observed for a transposable element family which is only present in two species, and indicates, particularly given the divergence time between the two species, that the family was present in both species prior to the transfer event. In other words, the horizontal transfer event did not introduce Helitron1_Dvir into a naïve genome, and the family can be assumed to have been present in the common ancestor of *D. virilis* and *D. mojavensis*. However, as the family is only present in two species, it is not possible to infer a likely direction of transfer, i.e. whether elements from *D. mojavensis* fall within the *D. virilis* clade or vice versa, as this would be dependent upon the position of a root, which cannot be inferred. Individual relationships between elements from *D. mojavensis* and those from *D. virilis* are well supported on both phylogenies.

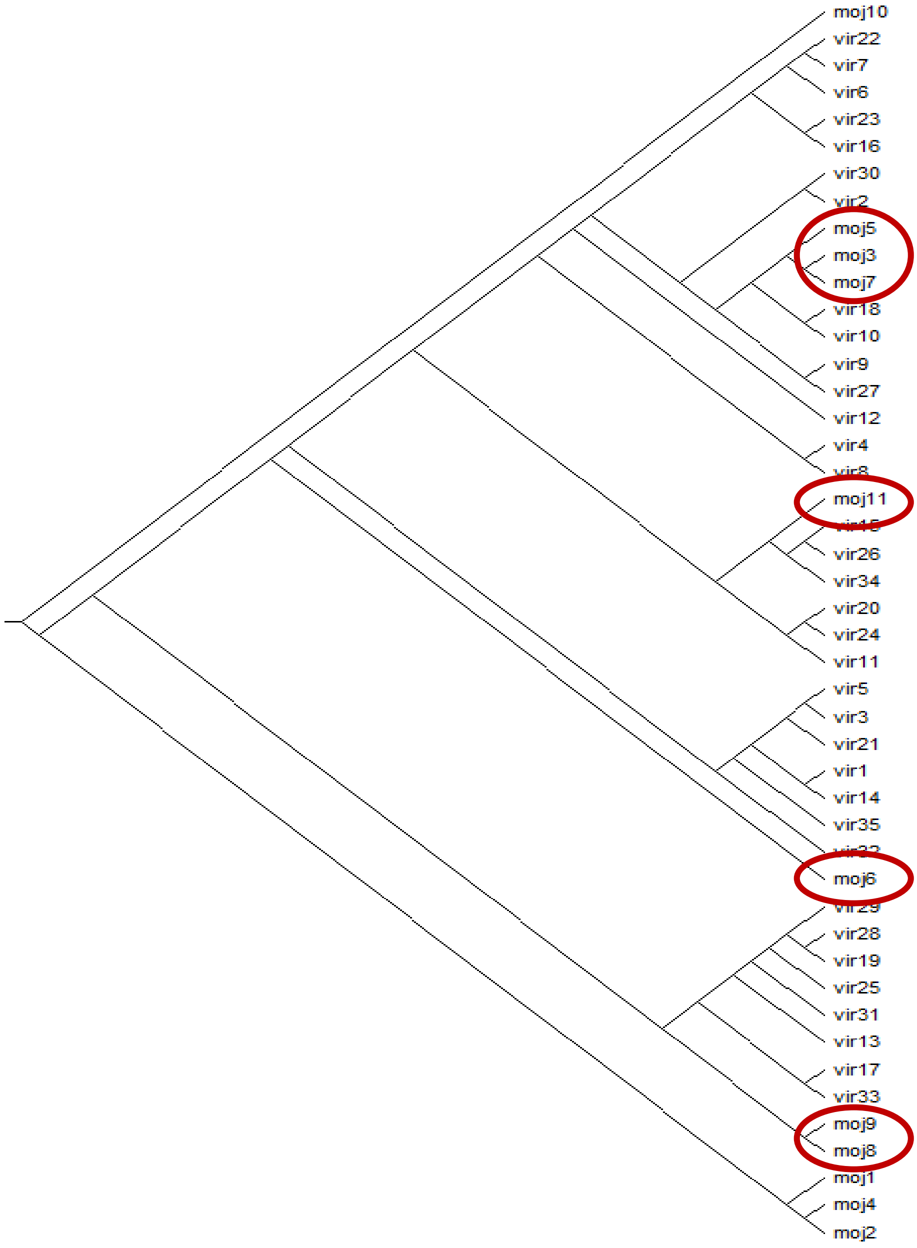


Figure 4.11: Maximum parsimony phylogeny of Helitron1_Dvir elements, demonstrating phylogenetic incongruence.

In the case of Helitron1_Dvir, although the smallest divergence observed is larger than the corresponding value for the *Adh* coding sequence, there is some evidence to suggest that horizontal transfer has occurred. It is possible that by requiring the stringent criterion of divergence between elements of the same family in different species being less than between the host genes, the extent of the influence of horizontal transfer on the evolution of transposable element sequences in *Drosophila* may be underestimated.

4.3.2.5 Effect of copy number

As suggested by the observations involving families present in both *D. persimilis* and *D. pseudoobscura*, it is possible that families which have a higher copy number of elements are more likely to have at least a single pair of elements which have diverged to a lesser extent than the host gene *Adh* than families with smaller copy numbers. This effect follows theoretical expectations, and will be discussed in chapter 5.

4.3.3 Non-LTR retrotransposons

The investigation of horizontal transfer in *Drosophila* was then extended to the non-LTR retrotransposons. In total, of the 56 families for which the consensus sequence was available on Repbase, 41 were present in at least two species out of the twelve *Drosophila* species for which the genome has been sequenced (Appendix 4). Examination of a broad range of non-LTR families was particularly interesting, as very few cases of horizontal transfer of non-LTR families have been reported. In a review of all reported cases of horizontal transfer in *Drosophila* (Loreto et al. 2008), only 5.0% of putative cases of horizontal transfer were attributed to non-LTR retrotransposons. Horizontal transfer of non-LTR retrotransposons therefore appears to be very rare, with only four families having been reported to have undergone this process: jockey, doc, F and I.

The following thirteen families yielded at least one interspecies comparison for which the smallest divergence obtained was less than that between the host *Adh* genes: BS2, doc, doc2, doc6, FW, G6, HelenaDS, hetA, LINEJ1, R2, TLD1, TLD2 and TLD3 (Table 4.4). Of these families, the majority are restricted to the Sophophora subgenus. Only two were distributed throughout both the Sophophora and Drosophila subgenera, and there were no families restricted to the Drosophila subgenus for which horizontal transfer was inferred. The BS2 and Helena_DS families are distributed throughout the *Drosophila* genus and appear to have been involved in horizontal transfer. Divergence data for the remaining non-LTR retrotransposon families, for which evidence for horizontal transfer was not found, are given in Appendix 4.

Family	Low divergence	Patchy distribution	Phylogenetic incongruence
BS2	Yes	Yes*	Yes
doc	Yes	Yes	Yes
doc2	Yes	No	Yes
doc6	Yes	Yes	Yes
G6	Yes	No	Yes
FW	Yes	No	No
Helena_DS	Yes	Yes	Yes
hetA	Yes	No	Yes
LINE J-1	Yes	Yes*	Yes
R2	Yes	Yes	Yes
TLD1	Yes	Yes	Yes
TLD2	Yes	Yes	Yes
TLD3	Yes	No	Yes

Table 4.4: Summary of the evidence for horizontal transfer among the non-LTR retrotransposons. An asterisk (*) indicates that patchy distribution is recorded as present although horizontal transfer to explain such a distribution is not supported by the divergence data.

Family	Estimated number of events	Species involved and inferred directions of transfer
BS2	2	yak>sim, mel<>sim
doc	4-5	mel>sim, yak>sim/sec, (sim<>sec), mel<>sec, mel<>yak
doc2	1	sim/sec>mel
doc6	6	sim<>sec, mel>yak, sim>(?)yak, sec>(?)yak, mel<>sim, mel<>ere
FW	1	(sim<>sec)
G6	2-3	(sim<>sec), mel<>sim, mel<>sec
HelenaDS	1	vir>ana
hetA	1	(sim<>sec)
LINE_J1	1-3	mel<>sim/sec, (sim<>sec)
R2	1	mel>ana
TLD1	3	(mel<>yak), (mel<>ere), (yak<>ere)
TLD2	1	(sim<>sec)
TLD3	1-2	mel<>sim/sec

Table 4.5: Summary of inferred horizontal transfers of non-LTR retrotransposons among the twelve *Drosophila* species. The estimated number of events given is a minimum number of events required to explain the observations made. Where multiple explanations are possible, the most parsimonious, i.e., that requiring the fewest horizontal transfer events, is presented. > indicates transfer from the first species to the second. <> indicates a transfer of unknown direction. ( ) indicates lack of certainty that a transfer event has occurred. >? indicates the direction of transfer is suggested, but is uncertain.

4.3.3.1 Non-LTR retrotransposons distributed throughout the *Drosophila* genus

The **BS2** family of non-LTR retrotransposons has an extensive distribution throughout the *Drosophila* phylogeny. It is found in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta* of the Sophophora, and *D. mojavensis* of the Drosophila subgenus. This patchy distribution across the phylogeny may suggest that horizontal transfer has introduced BS2 into *D. mojavensis* from one of the Sophophoran species in which it is present, or their ancestor. Comparison of BS2 elements in *D. melanogaster*, *D. simulans* and *D. yakuba* suggests that horizontal transfer of BS2 may have occurred within the Sophophora, however, divergence data do not suggest that horizontal transfer has occurred between the subgenera. Divergence smaller than the equivalent value for the *Adh* coding region was obtained for comparisons of *D. melanogaster* with *D. simulans* (0.007), and *D. simulans* with *D. yakuba* (0.026). Small divergence between *D. simulans* and *D. yakuba*, but not between *D. sechellia* and *D. yakuba* suggests that transfer may have occurred recently with *D. simulans* as the recipient. The phylogenies produced support both of these putative cases of horizontal transfer, as elements from *D. simulans* and *D. melanogaster* form close relationships on both phylogenies, and elements from *D. simulans* are found in the *D. yakuba* clade, supporting transfer involving *D. simulans* as the recipient species. Some precise relationships differ between the two phylogenetic construction methods, but these overall patterns are observed on both the maximum likelihood and maximum parsimony phylogenies.

Helena_DS is widely distributed across the *Drosophila* phylogeny. The family is absent only from three of the twelve species for which sequenced genomes are available: the two closely-related members of the *obscura* group, *D. pseudoobscura* and *D. persimilis*, and the representative of the Hawaiian *Drosophila*, *D. grimshawi*, as previously reported (Granzotto et al. 2009). Its wide distribution may suggest that Helena was present in the ancestor of the *Drosophila* genus and has been retained in most lineages (Granzotto et al. 2009). Its distribution could be explained by as little as two independent instances of stochastic loss. Alternatively, Helena may have achieved its distribution through horizontal transfer, possibly between the Sophophora and Drosophila subgenera. The only evidence for horizontal transfer of Helena obtained from interspecies comparisons comes from the comparison of *D. ananassae*, of the Sophophora, and *D. virilis*, of the Drosophila. In this case, the smallest divergence observed is smaller than that between the *Adh* genes, with a value of 0.235, compared with 0.252 for *Adh*. However, this distance is relatively large and suggests that if horizontal transfer of Helena has ever occurred, it was not a recent event. A relatively ancient horizontal transfer event between a species of the Sophophora and a species of the Drosophila subgenus would explain the current extensive distribution of Helena. Had the event occurred a long time in the past, it would not be expected to find small divergence between elements in closely-related species. Additionally, assuming that transposable elements are not under as great a level of constraint as host genes, such as *Adh*, the further back in time a horizontal transfer event has occurred, the less likely it would be to detect it. Mutations would accumulate rapidly and it would not take long for the divergence to exceed that of *Adh*. In this case, although *D. ananassae* and *D. virilis* provide the only interspecies comparison for which the divergence is smaller than *Adh*, there are other examples of comparisons between species of the Sophophora and Drosophila for which the divergence observed is smaller than might be expected given 40 million years of divergence and minimal, if any, constraint. The smallest divergence for each interspecies comparison involving one species of the Sophophora and one species of the Drosophila is shown in Table 4.6.

Interspecies comparison	**Smallest divergence**	***Adh* divergence**
mel/vir	0.264	0.237
mel/moj	0.271	0.220
sim/vir	0.256	0.231
sim/moj	0.235	0.216
sec/vir	0.248	0.231
sec/moj	0.229	0.216
yak/vir	0.257	0.237
yak/moj	0.240	0.220
ere/vir	0.302	0.231
ere/moj	0.267	0.220
ana/moj	0.239	0.226

Table 4.6: Smallest divergence observed between Helena elements for eleven interspecies comparisons between the Sophophora and Drosophila subgenera, compared with the equivalent divergence between the *Adh* coding regions.

These values suggest that any horizontal transfer event that did occur may have involved the ancestor of *D. virilis* and *D. mojavensis*, rather than *D. virilis* itself. This would date the transfer very far back in time, at over 24 million years ago. Alternatively, transfer may have occurred from *D. virilis* into the ancestor of the Sophophoran species in which Helena_DS is present, introducing the family to this lineage. Therefore, small divergence between these species and *D. mojavensis* would merely be a consequence of this transfer. However, under this scenario, divergence would be expected to be smaller than for *Adh* for all interspecies comparisons of Sophophoran species with *D. virilis*, rather than only *D. ananassae*. The phylogenies produced using Helena_DS elements from all species suggest that horizontal transfer of Helena_DS has occurred from *D. virilis* into *D. ananassae* more recently than expected, following the divergence of *D. virilis* and *D. mojavensis*, and of *D. ananassae* from its closest relatives for which the genome sequence is available. All elements from *D. ananassae* cluster together with all elements from *D. virilis*, just outside the clade of elements from *D. mojavensis* (Figure 4.12). This relationship is well-supported on both phylogenies. As all elements from *D. ananassae* cluster together with elements from *D. virilis*, this suggests that the transfer event may have introduced the family into *D. ananassae*. Alternatively, Helena_DS may have already been present in *D. ananassae*, which would explain the presence of the family in *D. melanogaster* and its relatives, and was lost prior to the transfer. Had Helena_DS elements been present in *D. ananassae* at the moment of transfer, it appears that only elements descended from the transfer have survived and are present in the contemporary *D. ananassae* genome.

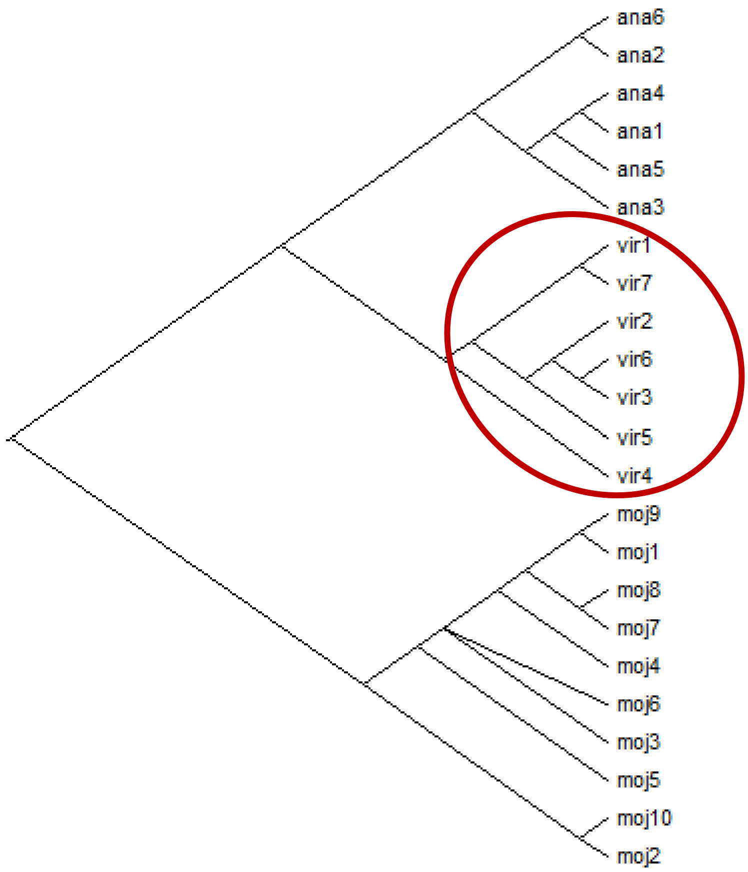


Figure 4.12: Section of the maximum parsimony phylogeny of Helena_DS elements, showing the incongruent position of elements from *D. virilis*.

4.3.3.2 Non-LTR retrotransposons restricted to the Sophophora

The remaining non-LTR retrotransposon families which are implicated to have been involved in horizontal transfer events are restricted to the Sophophora subgenus. Horizontal transfer of three of these families, LINE-J1, R2 and TLD1, is supported by all three lines of evidence investigated: small divergence between elements in different species, incongruence between the host and element phylogenies, and a patchy distribution across the host species tree.

LINE J-1 has an unusual distribution. It is found in the three closely-related species *D. melanogaster*, *D. simulans* and *D. sechellia*, and also in the more distantly related species *D. ananassae*, but is absent from *D. yakuba* and *D. erecta*. This distribution may suggest horizontal transfer between the ancestor of the three close relatives and *D. ananassae*, or may be explained by stochastic loss of LINE J-1 from the lineage leading to *D. yakuba* and *D. erecta*. The amount of divergence between elements in different species does not support the hypothesis of horizontal transfer involving *D. ananassae*. However, low levels of divergence between elements in *D. melanogaster*, *D. simulans* and *D. sechellia* suggest horizontal transfer of LINE J-1 may have occurred between these three species. Smallest divergence ranges from 0.003 between *D. simulans* and both *D. melanogaster* and *D. sechellia*, to 0.006 between *D. melanogaster* and *D. sechellia*. Such small divergence strongly suggests that horizontal transfer has occurred. Divergence between elements in these three species and *D. ananassae* is much greater than that observed between the *Adh* genes, and is consistent with vertical transmission, suggesting the family has indeed been lost from *D. yakuba* and *D. erecta*. The phylogenies produced using LINE J-1 elements are incongruent with the host phylogeny, but are inconsistent, most likely due to the extremely small divergence between many of the elements. Elements from *D. ananassae* group together on both trees, to the exclusion of elements from other species, therefore supporting the lack of involvement of this species in any horizontal transfer events. On the maximum parsimony tree, all elements from *D. melanogaster* cluster together, however, elements from *D. simulans* and *D. sechellia* do fall just outside this cluster, more closely related to the elements from *D. melanogaster* than to the others from *D. simulans* and *D. sechellia*. This may suggest transfer from *D. melanogaster* into the other two species. However, on the maximum likelihood phylogeny, elements from *D. melanogaster* are not found in single location on the tree, but are present in two separate locations, both within the *D. simulans*/*D. sechellia* clade, supporting transfer involving *D. melanogaster* as the recipient, the opposite of that suggested by the maximum parsimony tree. The branches leading to *D. melanogaster* elements are very poorly-supported on both phylogenies. Therefore, the LINE J-1 phylogenies cannot be used to confidently infer the potential direction of any transfer, although both do support transfer involving *D. melanogaster* and its closest relatives. Elements from *D. simulans* and *D. sechellia* are mixed together on both phylogenies, however, as discussed previously, this observation cannot provide convincing evidence for horizontal transfer between these species.

R2 is present in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. ananassae*, but is absent from *D. erecta*. This is likely to be due to stochastic loss of the family from this lineage. Alternatively, its patchy distribution may be explained by horizontal transfer, perhaps introducing the family into *D. ananassae*. For three interspecies comparisons, the smallest divergence observed is less than the equivalent value for the *Adh* coding region. These comparisons are between *D. ananassae* and the three closely-related species *D. melanogaster*, *D. simulans* and *D. sechellia*. This may suggest horizontal transfer has occurred in the past between the ancestor of these three close relatives and *D. ananassae*. However, the divergence data suggest a much more recent transfer involving *D. melanogaster* and *D. ananassae*, as the smallest divergence between these two species is only 0.007. Divergence of *D. simulans* and *D. sechellia* elements compared with those from *D. ananassae* are much larger, at 0.032 for *D. simulans* and 0.035 for *D. sechellia*, although these values are still considerably smaller than for the *Adh* gene. In addition, although slightly higher than the corresponding divergence for the *Adh* gene, the smallest divergence between R2 elements in *D. yakuba* and *D. ananassae* is 0.102, compared with 0.100 for *Adh*. The small divergence between elements from *D. simulans*, *D. sechellia* and *D. yakuba* when compared with *D. ananassae* is likely to be an artefact of their high degree of similarity with *D. melanogaster* elements, which is due to relatively recent common ancestry of these species. The divergence data therefore suggest that any transfer involved *D. melanogaster* as the donor species and *D. ananassae* as the recipient.

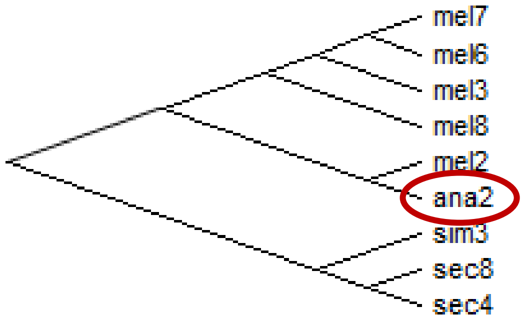


Figure 4.13: Section of the maximum parsimony phylogeny of R2 elements, showing the incongruent position of the element ana2 from *D. ananassae*.

This situation is comparable to that of the DNA transposon family Looper. The phylogenetic trees produced support this hypothesis. One of the *D. ananassae* elements, ana2, forms a close relationship with a *D. melanogaster* element, mel2, supported in an average of 99.2% of bootstrap replicates on the maximum parsimony phylogeny, and with a score of 0.975 on the maximum likelihood phylogeny. The ana2 element falls within a clade of *D. melanogaster* elements (Figure 4.13), which further supports the hypothesis of a transfer from *D. melanogaster* into *D. ananassae*, as does the presence of a second element from *D. ananassae* which does not fall in this position, and may potentially form the outgroup of R2 elements under vertical transmission, although it is not possible to root the tree. This element does not appear to be closely-related to any other element in the tree, and may resemble the R2 elements present in *D. ananassae* before the transfer from *D. melanogaster*. It is possible therefore that there are two “subfamilies” of R2 present in *D. ananassae*.

TLD1 is found only in *D. melanogaster*, *D. yakuba* and *D. erecta*. It is unusual for a transposable element family to be found in *D. melanogaster* and not its closest relatives *D. simulans* and *D. sechellia*. It is possible that the family has been lost stochastically from the lineage leading to these two species. Alternatively, TLD1 may have been involved in horizontal transfer. The interspecies comparisons of TLD1 elements support this hypothesis, as the smallest divergence observed for all three interspecies comparisons is 0.000, i.e. the sequences are identical. However, these results need to be interpreted with caution as the TLD1 consensus sequence is only 196bp in length. It is not unexpected to find unconstrained sequences in *D. melanogaster* and *D. yakuba* of such a short length with no mutations between them. Therefore, the high identity between elements in these species may simply be an artefact of the short length of the sequences, and not due to horizontal transfer.

The trees of the elements are incongruent with the host phylogeny. Although all *D. yakuba* elements consistently group together on both trees, an element from *D. erecta* clusters just outside of this group. Elements from *D. melanogaster* and *D. erecta* are found mixed together on the phylogenies. This is not surprising due to the high identity between sequences in the different species. Support for individual relationships is poor on almost every branch on both phylogenies. This is due to the extremely low level of variation between elements. In many cases, the topology cannot be resolved, and does not provide convincing evidence for horizontal transfer. Therefore despite the strikingly low divergence between TLD1 elements in different species, due to the short length of the TLD1 sequence and an inconclusive phylogeny, horizontal transfer of TLD1 elements cannot confidently be inferred to have occurred.

Horizontal transfer of the majority of non-LTR retrotransposons restricted to the Sophophora are supported by two lines of evidence. These are doc, doc2, doc6, FW, G6, hetA, and TLD3.

The **doc** family has been implicated in horizontal transfer previously (reviewed by Loreto et al. 2008). This family is contained to the *melanogaster* group, found in *D. melanogaster*, *D. simulans*, *D. sechellia* and *D. yakuba*. The absence of the family from *D. erecta* results in a patchy distribution across the phylogeny. All interspecies comparisons of doc elements yield divergence substantially smaller than for *Adh*, ranging from as little as 0.001 between *D. melanogaster* and *D. simulans*, to only 0.009 between *D. sechellia* and *D.*

yakuba. This extremely small level of divergence strongly supports the previously reported cases of horizontal transfer of the doc family. However, the range of divergence values for each interspecific comparison is quite large, suggesting that any horizontal transfer events have occurred between species in which the family is already present, rather than introducing the family into a naïve genome. The phylogenies produced suggest that multiple horizontal transfer events of doc elements may have occurred. Although some relatively large groups of elements from a single species form clades in the tree, many elements from different species are scattered throughout the phylogenies. Due to the extremely small divergence between elements in different species, and therefore the little information available to the phylogenetic reconstruction software, individual relationships between elements are inconsistent between the two tree-building methods. However, both phylogenies are incongruent and support the horizontal transfer cases suggested by the divergence data. As elements from various species are scattered throughout the trees, it is difficult to identify an underlying host phylogeny generated through vertical transmission, which can in most cases be determined. Therefore, it is difficult to confidently infer the direction of any cases of horizontal transfer. However, multiple examples of incongruence are consistent between the two trees. Elements from both *D. simulans* and *D. sechellia* are found within a cluster of elements from *D. yakuba*, possibly suggesting transfer from *D. yakuba* into *D. simulans*, *D. sechellia* or their ancestor. An element from *D. simulans* is consistently observed among a group of elements from *D. melanogaster*, suggesting transfer from *D. melanogaster* into *D. simulans*. A single element from *D. yakuba* is found in a cluster of elements from *D. melanogaster*, *D. simulans* and *D. sechellia*, and may have been involved in a transfer involving the ancestor of these species. Consistent incongruence involving *D. melanogaster* and *D. sechellia* is also observed. Further examples of incongruence, which are not consistent between the maximum parsimony and maximum likelihood trees, are also present. Many of the relationships presented on both phylogenies are well-supported, such as the clustering involving *D. yakuba* and *D. simulans*, present in an average of 98.8% of bootstrap replicates on the maximum parsimony tree, and with a score of 1.0 on the maximum likelihood tree. The trees, and the divergence data, give the impression of numerous horizontal transfer events involving doc. A series of recent transfers, involving transfer from a species which itself is a recipient of a previous transfer, may account for the small divergence between elements in different species without the requirement for so many transfers. However, the scrambled relationships consistently observed on the trees do suggest that many individual horizontal transfers of doc have occurred.

There are six families belonging to the doc group of non-LTR retrotransposons. Out of these six, in addition to doc, analyses conducted on a further two families (doc2 and doc6) also support horizontal transfer of these families. The remaining three doc families are not implicated to have been involved in horizontal transfer. The distribution of **doc2** does not support horizontal transfer, with the family present in, and restricted to, *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*. Unlike doc, interspecies comparisons of doc2 elements do not yield divergence smaller than between the *Adh* coding regions in all cases. Only comparisons of *D. melanogaster* with both *D. simulans* and *D. sechellia* are indicative of horizontal transfer, with smallest p distances of 0.018 and 0.017, respectively. These values are in line with the values for *Adh*, which are 0.019 and 0.017, respectively, rather than much smaller, as was the case for the doc family. Therefore, although there is some evidence from divergence to suggest horizontal transfer of doc2 may have occurred, alone it is not particularly convincing. However, the phylogenies produced are incongruent with known host relationships and provide further evidence to support the hypothesis of horizontal transfer. On both phylogenies, a single element from *D. melanogaster* is found within a large clade of 22 elements from *D. simulans* and *D. sechellia*, clustering with an element from *D. simulans*, suggesting transfer from *D. simulans*, or the ancestor of *D. simulans* and *D. sechellia,* into *D. melanogaster*. Elements from *D. melanogaster* are also found elsewhere on the tree, outside of the *D. simulans*/*D. sechellia* clade, suggesting that any such transfer did not introduce doc2 into *D. melanogaster*. On both phylogenies, incongruence is observed that supports horizontal transfers that are not suggested by the divergence data. For example, an element from *D. melanogaster* clusters with *D. erecta*. The position of elements from *D. yakuba* varies between the two phylogenies, and therefore elements in this species cannot confidently be inferred to have undergone horizontal transfer. Although phylogenetic incongruence is observed for doc2, low bootstrap values supporting the relationships which might be explained by horizontal transfer are found on the maximum parsimony phylogeny. However, some of these relationships do appear, well-supported, on the maximum likelihood phylogeny.

In common with doc, **doc6** elements are present in, and restricted to, *D. melanogaster*, *D. simulans*, *D. sechellia* and *D. yakuba*. Analysis of divergence between doc6 elements in different species indicates that horizontal transfer may have occurred. The smallest divergence observed for comparisons of elements from *D. melanogaster*, *D. simulans*, *D. sechellia* and *D. yakuba* with each of the other three species are all smaller than the corresponding values for *Adh*, and, as in the case of doc, are strikingly small.

Identical elements are found in *D. simulans* and *D. sechellia*, suggesting transfer of doc6 may have occurred between these two closely-related species. Divergence can be as small as 0.001 in comparisons of *D. melanogaster* with both *D. simulans* and *D. sechellia*, and between *D. simulans* and *D. yakuba*. Smallest divergence of 0.002 is observed for the comparison of elements from *D. sechellia* with *D. yakuba*. These distances represent extremely low levels of divergence which are unlikely to be explained in the absence of horizontal transfer. However, for such low values to be obtained for all comparisons, given the divergence times between the host species, multiple transfer events would be required. In the phylogenies constructed using doc6 sequences, as in most cases where an element family is present in *D. simulans* and *D. sechellia*, the elements from these species are found amongst each other in the phylogeny. This is not considered to be particularly informative. However, further incongruence is observed between the phylogenies constructed and known host relationships. The two trees are, however, generally inconsistent, most likely as a result of the minimal amount of mutation data available to reconstruct relationships. There is some consistency between the two trees which does support at least two horizontal transfer events. On the maximum likelihood phylogeny, two elements from *D. yakuba* cluster with an element from *D. melanogaster* in a clade of elements from *D. melanogaster*. This relationship is also observed on the maximum parsimony tree, however, these elements fall just outside the *D. melanogaster* clade. This probably indicates horizontal transfer from *D. melanogaster* into *D. yakuba*. Close relationships between elements from *D. simulans* and *D. yakuba* are observed on both phylogenies, and on the maximum parsimony tree, close relationships between *D. yakuba* and *D. sechellia* elements are also present. In both cases, the elements from *D. yakuba* involved in these relationships are found within the *D. simulans*/*D. sechellia* clade, perhaps suggesting horizontal transfer has occurred from *D. simulans* into *D. yakuba*, as indicated by the maximum likelihood tree. It is difficult to confidently determine the potential direction of transfer as elements from each species are scattered throughout the trees, and the phylogenies are generally inconsistent.

FW exhibits a distribution across the *Drosophila* phylogeny consistent with vertical transmission, present in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*. The family represents a relatively weak candidate for horizontal transfer, with divergence between elements smaller than the host *Adh* coding region for only one interspecies comparison, between the closely-related species *D. simulans* and *D. sechellia*. Divergence between FW elements in these two species can be extremely small, at as little as 0.001, however, given the very close relationship between these host species, it may be possible for such small divergence to occur by chance. The phylogenies produced do indeed show a mixed clade of elements from *D. simulans* and *D. sechellia*, however, as discussed previously, this cannot be accepted as convincing evidence for horizontal transfer between these two host species. FW therefore represents a possible case of horizontal transfer, but is not supported by extensive evidence.

G6 is found in *D. melanogaster*, *D. simulans* and *D. sechellia*. Its distribution is therefore very restricted and does not suggest that horizontal transfer has occurred. However, interspecies comparisons of G6 elements reveal smallest divergence less than that between the *Adh* coding regions for all three possible pairwise comparisons. As in the case of doc and doc6, these are all strikingly small and suggest very recent horizontal transfer, ranging from 0.001 between *D. melanogaster* and *D. sechellia*, to 0.005 between *D.*

melanogaster and *D. simulans*. These values are too small to be explained by a single transfer event involving *D. melanogaster* and the ancestor of *D. simulans* and *D. sechellia*, therefore at least two horizontal transfer events can be inferred. Once again, the range of divergence values for all interspecies comparisons is quite large, suggesting any transfer events did not introduce G6 into a naïve genome, but instead have occurred between species in which a population of G6 elements was already established. Close relationships between elements from *D. simulans* and *D. melanogaster*, and *D. sechellia* and *D. melanogaster* are consistently observed on both phylogenies, although some precise relationships differ between the two phylogenetic construction methods. This is probably accounted for by the extremely small divergence between some of the elements. Elements from *D. melanogaster* cluster with both *D. simulans* and *D. sechellia* in different parts of the *D. simulans*/*D. sechellia* clade, supporting at least two separate transfers occurring since the divergence of these two closely related species. The phylogeny of G6 elements is clearly incongruent with the phylogeny of the host species, with high levels of support for branches grouping elements from different species together. This is indicative of horizontal transfer of G6 among *D. melanogaster*, *D. simulans* and *D. sechellia*. It appears from the phylogeny that G6 was already present in each of the species before horizontal transfer occurred, rather than being introduced into one of the species by this method. This is due to the presence of individual elements clustering with those from other species, rather than all the elements from one species clustering together in an unexpected position on the tree.

The **hetA** family is found only in the closely-related species *D. melanogaster*, *D. simulans* and *D. sechellia*. Its distribution is therefore not indicative of horizontal transfer. In addition, the *Drosophila* 12 Genomes Consortium reported that hetA appears to have followed a pattern of vertical transmission (Clark et al. 2007). The pairwise comparisons of hetA elements from *D. melanogaster* with those from the other two species do not yield divergence smaller than that observed between the host *Adh* coding regions. However, the smallest divergence between an element from *D. simulans* and one from *D. sechellia* is very small, at only 0.001. This small amount of divergence could be due to horizontal transfer between these two species. However, *D. simulans* and *D. sechellia* are extremely closely related, and the high identity between elements could simply be due to chance. The phylogenetic trees of hetA do indeed reveal elements from *D. simulans* and *D. sechellia* mixed amongst each other. However, as explained previously, phylogenetic trees of elements found in very closely-related species, such as *D. simulans*/*D. sechellia* or *D. pseudoobscura*/*D. persimilis*, tend to form mixed clades of elements from the two closely-related species, regardless of whether or not horizontal transfer is inferred. Therefore, the observation of phylogenetic incongruence between *D. simulans* and *D. sechellia* in the case of hetA elements cannot be taken as convincing evidence that horizontal transfer between these two species has occurred. Interestingly, however, the individual relationships between elements from *D. simulans* and *D. sechellia* are consistent between the two phylogenetic construction methods, and are very well-supported in both cases. There are three groupings of elements from *D. simulans* and *D. sechellia* present on both phylogenies, supported in between 72.8% and 100% of bootstrap replicates on the maximum parsimony tree, and with a score of 0.887 to 1.0 on the maximum likelihood tree.

TLD3 is found in *D. melanogaster*, *D. simulans* and *D. sechellia*. Its distribution is therefore, as in many of the other putative cases of horizontal transfer of non-LTR retrotransposons, unsupportive of the hypothesis of horizontal transfer. The smallest divergence observed, however, for both interspecies comparisons involving *D. melanogaster*, is less than that observed between the host *Adh* coding regions, at 0.000, i.e. identical elements in *D. melanogaster* and *D. sechellia*, and 0.006 in the comparison of *D. melanogaster* and *D. simulans*. As with the other TLD families, TLD3 is represented by a very short consensus sequence of 172bp. The small divergence values obtained may therefore simply be due to chance, as a result of the short sequence length and the relatively close relationships between the host species. The phylogenies produced are poorly resolved, with many elements branching off from a single point, as many elements, particularly in *D. melanogaster*, are identical to each other, so individual relationships cannot be determined. This has generated star-like phylogenies which are indeed incongruent with known host relationships. Elements from all three species form close relationships supported by both phylogenetic construction methods. However, due to the poorly-resolved relationships in the phylogenies, they do not provide any substantial further support than that provided from examination of the divergence between elements. Given the short length of TLD3 elements, this small divergence alone does not provide convincing evidence for horizontal transfer of this family between *D. melanogaster* and its closest relatives.

Horizontal transfer of only a single non-LTR retrotransposon family, **TLD2**, is supported by only small divergence between elements in different species. Although phylogenetic incongruence is observed for this family, it is only between the closely related species *D. simulans* and *D. sechellia*, and therefore cannot be considered to be reliable. TLD2 has a more extensive distribution than TLD1, found in *D. melanogaster*, *D. simulans*, *D. sechellia* and *D. yakuba*. Its distribution, and the majority of interspecies comparisons, do not suggest that TLD2 has been horizontally transferred among these species. Divergence smaller than the corresponding value for *Adh* is obtained for the comparison of *D. simulans* and *D. sechellia* elements, with a value of 0.000. As in the case of TLD1, this has to be treated with caution. The consensus sequence of TLD2 is only 218bp in length, and therefore is likely to be invariant between species simply due to its short length. Additionally, *D. simulans* and *D. sechellia* are extremely closely related, having diverged only around 2 million years ago. This makes it much more likely that this identity between elements in the two species is due to chance rather than horizontal transfer, even more so than for TLD1. As was the case for hetA, for which, like TLD2, a divergence smaller than *Adh* was obtained for the comparison of *D. simulans* and *D. sechellia* elements only, production of a phylogeny is unlikely to provide any convincing evidence for whether or not horizontal transfer of the family has occurred between these two species. This is indeed the case for TLD2. All elements from *D. melanogaster* form a well-supported monophyletic clade, as do all elements from *D. yakuba*. As expected, elements from *D. simulans* and *D. sechellia* form a clade on the tree, with the elements from the two species mixed amongst each other. Within this clade, the majority of branches are poorly supported, although some specific close relationships between elements from *D. simulans* and *D. sechellia* are consistently supported on both phylogenies, perhaps suggesting horizontal transfer between these two species has occurred. However, due to the short sequence length and the close relationship between the two host species, these observations cannot be taken as convincing evidence for horizontal transfer of TLD2 between *D. simulans* and *D. sechellia*.

4.3.3.3 Summary of horizontal transfer of non-LTR retrotransposons

Interestingly, although the thirteen families described above may potentially have been involved in horizontal transfer events, which is more than would be expected from previous estimates, many of these families are of the same type. BS2, doc, doc2, doc6, G6 and Helena all belong to the Jockey superfamily. TLD1, TLD2, TLD3 and hetA are all telomeric sequences. R2 is the only family described above which does not fall into either of these two groups, and is classified in the R2 superfamily.

Analysis of thirteen families of non-LTR retrotransposons out of 41 reveals evidence that horizontal transfer has occurred. This is a much greater number than expected from previous observations (Loreto et al. 2008). However, in many cases, the evidence supporting the hypothesis of horizontal transfer is relatively poor, for example, the presence of phylogenetic incongruence supported in a small number of replicates. The most convincing cases are those for which the divergence between elements is much smaller than can be expected.

4.3.3.4 Transcriptional readthrough

Due to their mechanism of retrotransposition, non-LTR retrotransposons have been reported to have undergone transcriptional readthrough past the terminator, and therefore cause the mobilisation of adjacent host DNA during transposition. This concept is particularly interesting to investigate with respect to horizontal transfer, as it might indicate a potential mechanism through which host DNA could be transferred between *Drosophila* species. This has implications for the application of transposable elements to transgenesis, whereby the mobilising activity of transposable elements is exploited to introduce DNA into a foreign genome. The use of transposable elements in this process makes the assumption that the foreign DNA that is inserted is unable to leave the system and infect another species. Beyond transgenesis, this process has potential implications for the movement of DNA in natural systems, and may represent a means by which genomes can acquire fragments of novel DNA sequence, which might potentially be exapted to perform new functions.

For each of the families for which the smallest p distance obtained was less than that for the *Adh* gene, i.e. those that could potentially be inferred to have undergone horizontal transfer, this phenomenon was investigated. 3’ flanking host DNA was extracted for each element for which a comparison with an element of the same family from another species yielded divergence less than for *Adh*. These are the pairs of elements which may potentially share high sequence identity due to horizontal transfer. If this is indeed the case, and transcriptional readthrough has occurred, the pair of elements would be expected to share sequence identity in the flanking DNA downstream from the 3’ end of the element. However, no evidence of this process was found. In all cases, the 3’ flanking host DNA present was unique to each individual element.

4.3.4 LTR Retrotransposons

Finally, the investigation of horizontal transfer in *Drosophila* was extended to the LTR retrotransposons. In total, 64 families were examined. Out of these, 59 were found to be present in at least two of the *Drosophila* species for which the complete genome sequence is available (Appendix 5).

LTR retrotransposons provide an additional source of information when investigating their evolution when compared with non-LTR retrotransposons and DNA transposons: the presence of LTRs. Upon integration of an LTR retrotransposon into the genome, its 5’ and 3’ flanking LTRs are identical to each other. This is due to their mechanism of retrotransposition. Only the sequence between the two repeated regions in the flanking LTRs is transcribed. Upon integration, the U3 region from one end, and the U5 region from the other, is copied on to the ends of the element, next to each R region, regenerating a complete element. Consequently, the flanking LTRs are identical at the moment of integration, but will accumulate mutations and diverge from each other over time. As a result of this, the amount of divergence between the flanking LTRs can be used as a measure of the age of the insertion. Assessment of the amount of divergence among elements can give an indication of the age of the family in that species. However, as transposable elements are rapidly deleted from *Drosophila* genomes, it is unlikely that older elements would still be present. Therefore, this estimation may not provide a good indication of age, but can suggest whether or not a family is currently actively retrotransposing in a particular species, and if not, how recently its proliferation stopped. Additionally, when examining a transposable element family as a whole, solo LTRs can be more prevalent than the corresponding internal portions of the element, although in some individual species this is not the case. This, in addition to the availability of two flanking LTRs for each full-length insertion, provides more data to be examined. In the case of four LTR retrotransposon families, solo LTRs are found in a species in which the full-length element, or even fragments of the internal sequence, are no longer found (Table 4.7). This confirms that the absence of the family from this species is due to stochastic loss, rather than the element having never been present. This allows an additional species to be incorporated into the evolutionary analyses. Phylogenetic trees can be constructed for LTR sequences as well as full-length elements or internal sequences. The use of solo LTRs, especially in species in which the remainder of the element is not found, gives these phylogenetic analyses more power to resolve relationships. These trees might also provide evidence for additional transfer events which cannot be detected by examination of internal sequences, if elements descended from the transfer event have been lost by recombination. However, phylogenies produced using solo LTR sequences are not as reliable as those produced using internal sequences, due to their short length and high copy number. Both of these factors may result in similarity between LTRs due to random chance rather than recent common ancestry.

Family	**Species**
Gypsy8	*D. sechellia*
Quasimodo	*D. simulans*, *D. sechellia*
Stalker2	*D. yakuba*, *D. erecta*
Tabor	*D. yakuba*

Table 4.7: Species in which solo LTRs of four LTR retrotransposon families are observed in the absence of the associated internal sequences. The loss of Gypsy8 from *D. sechellia* is presumably recent, as internal sequences are still observed in *D. simulans*.

The vast majority of LTR retrotransposons analysed (45/59) are restricted to the Sophophora. A further thirteen families are distributed across both the Sophophora and Drosophila subgenera. Only one family, TV1, is found in two species of the Drosophila subgenus and absent from the nine sequenced Sophophora genomes. This can be explained by the relative abundance of the Sophophora in the twelve sequenced genomes (75%), and the fact that most consensus sequences available for transposable element families correspond to elements found in Sophophoran species.

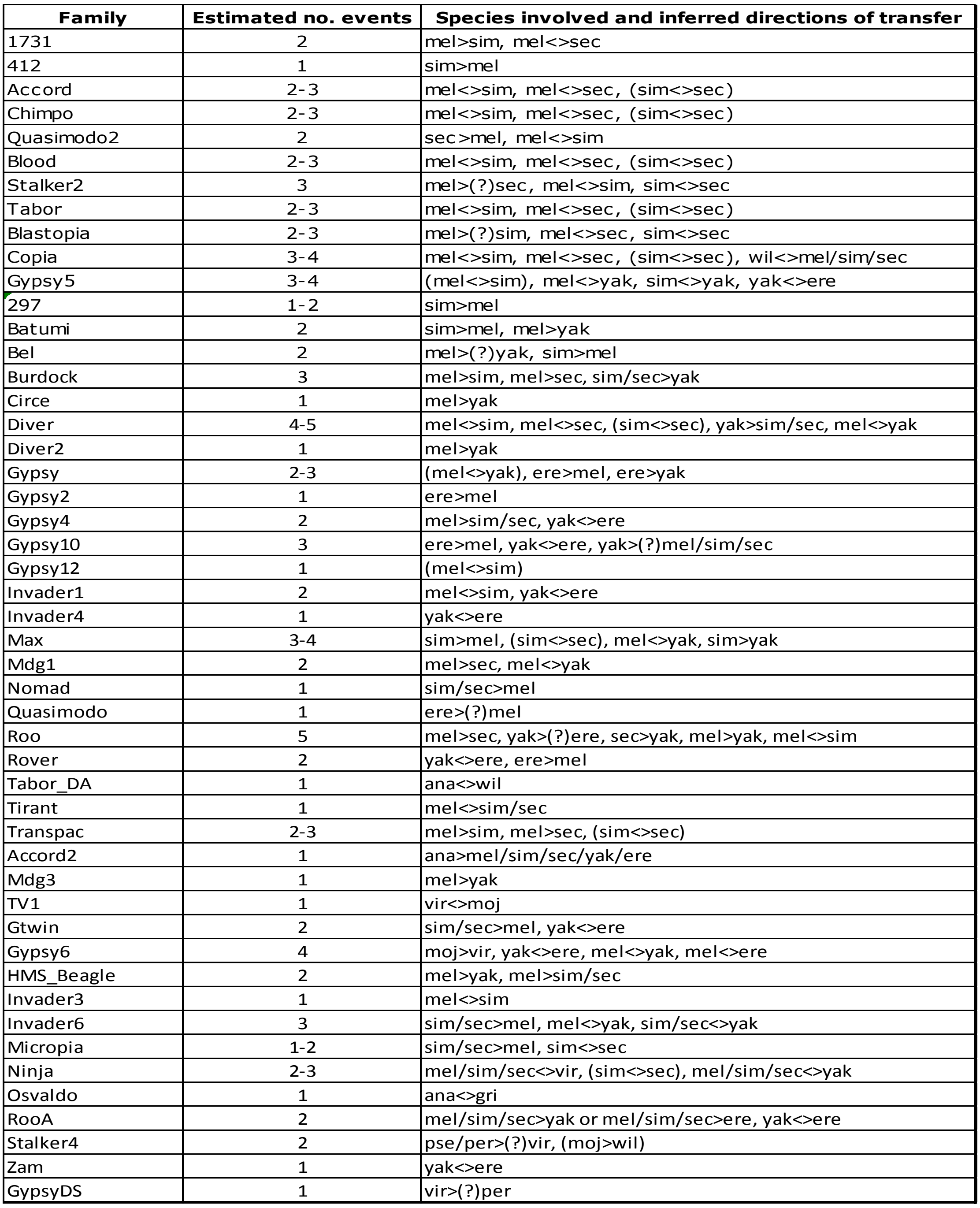
4.3.4.1 Summary of results for LTR retrotransposons

49 out of the 59 families of LTR retrotransposons which were present in at least two species yielded at least one interspecies comparison for which the smallest p distance obtained was less than that between the host *Adh* genes. This suggests, contrary to expectations, that horizontal transfer is more common for LTR retrotransposons than the other two types of transposable element. Following completion of this study, this result was confirmed by another study (Bartolome et al. 2009). Many of these putative cases of horizontal transfer are further supported by phylogenetic incongruence and a patchy distribution across the host species phylogeny (Table 4.8). Divergence data for the remaining LTR retrotransposon families, for which evidence for horizontal transfer was not found, are given in Appendix 5.

Family	Phylogenetic incongruence	Patchy distribution
1731	Yes	No
297	Yes	No
412	Yes	No
Accord	Yes*	No
Accord2	No	No
Batumi	Yes*	Yes
Bel	Yes	Yes
Blastopia	Yes	Yes
Blood	No	No
Burdock	Yes	No
Chimpo	Yes*	No
Circe	Yes	Yes
Copia	Yes	Yes
Diver	Yes*	Yes
Diver2	Yes	Yes
Gtwin	Yes	Yes
Gypsy	Yes	No
GypsyDS	No	Yes

Family	Phylogenetic incongruence	Patchy distribution
Gypsy2	Yes	No
Gypsy4	Yes	No
Gypsy5	Yes	Yes
Gypsy6	Yes	Yes
Gypsy10	Yes	No
Gypsy12	No	Yes
HMS Beagle	Yes	Yes
Invader1	Yes	No
Invader3	Yes	Yes
Invader4	Yes	No
Invader6	Yes	Yes
Max	Yes	Yes
Mdg1	Yes	No
Mdg3	No	No
Micropia	Yes*	Yes
Ninja	Yes	Yes
Nomad	Yes*	No
Osvaldo	Yes	Yes
Quasimodo	No	Yes
Quasimodo2	Yes	No
Roo	Yes	No
RooA	Yes	Yes
Rover	Yes	No
Stalker2	No	No
Stalker4	Yes	Yes
Tabor	No	No
Tabor_DA	No	Yes
Tirant	Yes	No
Transpac	Yes*	No
TV1	No	No
Zam	Yes	Yes

Table 4.8: Summary of the evidence for horizontal transfer of LTR retrotransposons in *Drosophila*. An asterisk (*) indicates that although both maximum parsimony and maximum likelihood phylogenies demonstrate incongruence with the host phylogeny supporting the hypothesis of horizontal transfer (i.e. incongruencies involving the species for which p<*Adh*), no specific incongruent relationships are consistent between the two phylogenetic construction methods.



Family	Estimated no. events	Species involved and inferred directions of transfer
1731	2	mel>sim, mel<>sec
412	1	sim>mel
Accord	2-3	mel<>sim, mel<>sec, (sim<>sec)
Chimpo	2-3	mel<>sim, mel<>sec, (sim<>sec)
Quasimodo2	2	sec>mel, mel<>sim
Blood	2-3	mel<>sim, mel<>sec, (sim<>sec)
Stalker2	3	mel>(?)sec, mel<>sim, sim<>sec
Tabor	2-3	mel<>sim, mel<>sec, (sim<>sec)
Blastopia	2-3	mel>(?)sim, mel<>sec, sim<>sec
Copia	3-4	mel<>sim, mel<>sec, (sim<>sec), wil<>mel/sim/sec
Gypsy5	3-4	(mel<>sim), mel<>yak, sim<>yak, yak<>ere
297	1-2	sim>mel
Batumi	2	sim>mel, mel>yak
Bel	2	mel>(?)yak, sim>mel
Burdock	3	mel>sim, mel>sec, sim/sec>yak
Circe	1	mel>yak
Diver	4-5	mel<>sim, mel<>sec, (sim<>sec), yak>sim/sec, mel<>yak
Diver2	1	mel>yak
Gypsy	2-3	(mel<>yak), ere>mel, ere>yak
Gypsy2	1	ere>mel
Gypsy4	2	mel>sim/sec, yak<>ere
Gypsy10	3	ere>mel, yak<>ere, yak>(?)mel/sim/sec
Gypsy12	1	(mel<>sim)
Invader1	2	mel<>sim, yak<>ere
Invader4	1	yak<>ere
Max	3-4	sim>mel, (sim<>sec), mel<>yak, sim>yak
Mdg1	2	mel>sec, mel<>yak
Nomad	1	sim/sec>mel
Quasimodo	1	ere>(?)mel
Roo	5	mel>sec, yak>(?)ere, sec>yak, mel>yak, mel<>sim
Rover	2	yak<>ere, ere>mel
Tabor_DA	1	ana<>wil
Tirant	1	mel<>sim/sec
Transpac	2-3	mel>sim, mel>sec, (sim<>sec)
Accord2	1	ana>mel/sim/sec/yak/ere
Mdg3	1	mel>yak
TV1	1	vir<>moj
Gtwin	2	sim/sec>mel, yak<>ere
Gypsy6	4	moj>vir, yak<>ere, mel<>yak, mel<>ere
HMS_Beagle	2	mel>yak, mel>sim/sec
Invader3	1	mel<>sim
Invader6	3	sim/sec>mel, mel<>yak, sim/sec<>yak
Micropia	1-2	sim/sec>mel, sim<>sec
Ninja	2-3	mel/sim/sec<>vir, (sim<>sec), mel/sim/sec<>yak
Osvaldo	1	ana<>gri
RooA	2	mel/sim/sec>yak or mel/sim/sec>ere, yak<>ere
Stalker4	2	pse/per>(?)vir, (moj>wil)
Zam	1	yak<>ere
GypsyDS	1	vir>(?)per

Table 4.9: Summary of inferred horizontal transfers of LTR retrotransposons among the twelve *Drosophila* species. The estimated number of events given is a minimum number of events required to explain the observations made. Where multiple explanations are possible, the most parsimonious, i.e., that requiring the fewest horizontal transfer events, is presented. > indicates transfer from the first species to the second. <> indicates a transfer of unknown direction. ( ) indicates lack of certainty that a transfer event has occurred. >? indicates the direction of transfer is suggested, but is uncertain.

4.3.4.2 LTR retrotransposons restricted to *D. melanogaster*, *D. simulans* and *D. sechellia*

All of the families discussed below are present in all three species (*D. melanogaster*, *D. simulans* and *D. sechellia*), but are restricted to these species. Consequently, patchy distribution cannot be observed for these families. However, the case for horizontal transfer for five families is supported by both small divergence and phylogenetic incongruence. These families are 1731, 412, Accord, Chimpo and Quasimodo2.

Horizontal transfer of family **1731** appears to have occurred between *D. melanogaster* and the other two species, but not between *D. simulans* and *D. sechellia*. Divergence between elements can be extremely small, with a smallest p distance of 0.003 between *D. melanogaster* and *D. simulans*, and 0.007 between *D. melanogaster* and *D. sechellia*. The range of divergence is, however, quite large, suggesting elements transmitted vertically were resident in each of the three species before the proposed recent transfers occurred, rather than the family being introduced into one or more species by horizontal transfer. The phylogenies of 1731 elements support the hypothesis of horizontal transfer involving *D. simulans* (Figure 4.14), as a single element from *D. simulans* falls within the *D. melanogaster* clade. However, the elements from *D. sechellia* cluster with the remaining elements from *D. simulans*, to the exclusion of all elements from *D. melanogaster*. The phylogenies therefore do not provide further evidence to support a transfer involving *D. melanogaster* and *D. sechellia*.

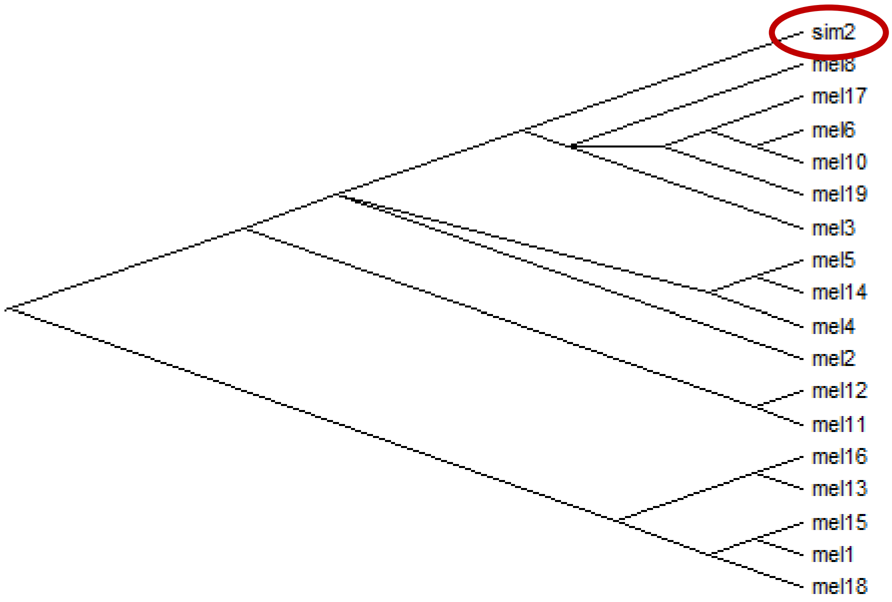


Figure 4.14: Section of the maximum parsimony phylogeny of 1731 elements, showing the incongruent position of one element from *D. simulans* within the clade of elements from *D. melanogaster*.

412 is a striking candidate for horizontal transfer, with a smallest divergence between elements in *D. melanogaster* and *D. simulans* of 0.001. In addition to being considerably smaller than the divergence between the host gene *Adh* for these species, the divergence is also smaller than that between elements in the closely-related species *D. simulans* and *D. sechellia*. Divergence is also smaller than for *Adh* when comparing elements from *D. melanogaster* and *D. sechellia*, with a smallest divergence of 0.009. The element in *D. simulans* which shares high identity with elements from *D. melanogaster*, designated sim1, appears to be a relatively recent insertion, with only three mutations between its flanking LTRs. Interestingly, there are many further solo LTRs in *D. simulans* which resemble these LTRs, although LTRs of this type are not found in *D. melanogaster*. This might suggest the transfer between these species was not very recent, although clearly the identity between elements suggests the contrary.

There is a polymorphic 63bp insertion or deletion between positions 609 and 671 in *D. melanogaster* and *D. sechellia*, with the full-length form present in *D. simulans*, which may be used to determine relationships between sequences. This would require the assumption that the event leading to this mutation had only occurred once, however, the insertion or deletion event appears to have occurred multiple times in parallel due to tandem duplication or replication slippage. Parallel deletion is a more likely explanation as there is only one element, in *D. sechellia*, for which the two tandem repeats of this sequence are identical. Therefore this mutation cannot be reliably used to infer relationships between sequences.

It has previously been reported that 412 elements have undergone horizontal transfer events between *Drosophila melanogaster* and *D. simulans* (Sanchez-Gracia et al. 2005). This supposition is based on very low divergence between *D. melanogaster* and *D. simulans*, and also on the incongruence between the species tree and a tree of 412 sequences, generated using the neighbour-joining method with bootstrapping. However, neighbour joining may not be an appropriate method to use for sequences with such low divergence. This study was also performed using a limited set of 412 sequences, obtained using PCR. However, the phylogenetic trees of the complete set of 412 elements from the genome sequences of *D. melanogaster*, *D. simulans* and *D. sechellia* constructed using maximum parsimony and maximum likelihood are also incongruent with the host phylogeny, supporting the conclusions drawn from the neighbour joining tree. Some precise relationships differ between the two methods, however, a cluster of six *D. melanogaster* elements is consistently found within a clade of *D. simulans* elements. This may suggest that transfer occurred from *D. simulans* into *D. melanogaster*.

Accord is an interesting transposable element family, as it appears to have been exapted to perform a function in *Drosophila*. There is a functional interaction between Accord and host gene *cyp6*g1 found upstream, which is implicated in DDT resistance (Chung et al. 2007). It is therefore tempting to speculate that *Drosophila* species have made use of sequence available to them, in the form of an Accord element, to improve resistance to DDT. More broadly, this could suggest that a strong enough selection pressure may lead species to utilise transposable element sequences, as a possible source of novel sequence information, which would have an interesting impact on transposable element evolution. Exaptation of transposable element sequences is contrary to the general pattern in *Drosophila*, by which transposable elements are deleterious and are rapidly deleted from the genome. Accord is present in *D. melanogaster*, *D. simulans* and *D. sechellia*, however there are no full-length elements in *D. simulans,* and large internal deletions appear to be common. Divergence between elements in different species is extremely small, and strongly supports the hypothesis of recent horizontal transfer. The smallest divergence between elements in *D. sechellia* and both *D. melanogaster* and *D. simulans* is 0.000, with a maximum divergence of 0.008 and 0.016, respectively. Smallest divergence between *D. melanogaster* and *D. simulans* elements is 0.001, with a maximum divergence of 0.009. Horizontal transfer is further supported by phylogenetic incongruence. The maximum parsimony and maximum likelihood trees are both incongruent, but generally inconsistent, most likely to due to the extremely small amount of mutation data available for these sequences. Both trees support a close relationship between the elements designated sec7 and mel8, however, in one tree this relationship is found within a clade of *D. sechellia* elements, and in the other, within a clade of *D. melanogaster* elements.

Divergence smaller than for the host gene *Adh* provides evidence for the horizontal transfer of **Chimpo** for all three interspecies comparisons. Divergence ranges from 0.006 to 0.071, 0.006 to 0.054, and 0.003 to 0.066 for the *D. melanogaster-D. simulans*, *D. melanogaster-D. sechellia*, and *D. simulans-D. sechellia* comparisons, respectively. The phylogenetic trees produced using the two methods are both incongruent with the host phylogeny, but precise relationships are inconsistent, most likely, as in the case of Accord, due to the small amount of divergence between many of the elements. Elements from all three species fall in various parts of the tree, but the relationships observed are unreliable.

The final LTR retrotransposon family found in *D. melanogaster*, *D. simulans* and *D. sechellia* only, for which horizontal transfer is supported by both small divergence and phylogenetic incongruence, is **Quasimodo2**. As in the case of 1731, there is no evidence to suggest transfer of this family between *D. simulans* and *D. sechellia*, but the remaining two transfer routes are supported by divergence smaller than *Adh* of 0.005 for the *D. melanogaster-D. simulans* comparison, and 0.011 for the *D. melanogaster-D. sechellia* comparison. However, values are not consistently small, and again it appears that transfer may have occurred into genomes that already contained a resident population of Quasimodo2 elements. A single element from *D. melanogaster* falls within the *D. sechellia* clade on both the maximum parsimony and maximum likelihood trees, supporting transfer from *D. sechellia* into *D. melanogaster*. The phylogenies provide no evidence for transfer involving *D. simulans*, however, it is possible that the small divergence observed here is an artefact of the transfer involving *D. sechellia*. As *D. simulans* and *D. sechellia* are extremely closely related, the *D. simulans* elements which share high identity with elements in *D. sechellia* related to the element involved in the transfer to *D. melanogaster*, would now be observed as sharing high identity with elements in *D. melanogaster* in the absence of direct horizontal transfer between these two species.

Horizontal transfer of several families restricted to *D. melanogaster*, *D. simulans* and *D. sechellia*, Blood, Stalker2 and Tabor, is supported by a single piece of evidence. That is, the divergence between elements found in different species is smaller than that for the host gene *Adh*.

Blood is implicated as being involved in horizontal transfer due to small divergence between elements for all three interspecies comparisons. Smallest divergence is extremely low, at 0.005 in comparison of *D. melanogaster* and *D. simulans* elements, 0.009 for *D. melanogaster* and *D. sechellia*, and 0.003 for *D. simulans* and *D. sechellia*. Although these values are extremely small, there are many divergent elements also present, suggesting that if any recent horizontal transfer has occurred, this did not introduce the family to any of these species, but instead introduced new elements to an existing population of Blood elements. However, the phylogenies produced do not support this hypothesis, as they are congruent with the host phylogeny, with the exception of *D. simulans* and *D. sechellia*, which, as described previously, can be ignored.

Stalker2 is an interesting case to investigate as although its distribution is restricted to *D. melanogaster*, *D. simulans* and *D. sechellia*, solo LTRs of this family are found in *D. yakuba* and *D. erecta*. This indicates that the Stalker2 family has been stochastically lost from the lineage leading to these species, and was present in the common ancestor of the five species. The sequences of the solo LTRs, although not as reliable to study due to their short sequence length and high copy number, which can make genuine relationships between sequences more difficult to detect, can be used to potentially infer whether any horizontal transfer occurred within the two species from which Stalker2 is now absent. Both the current distribution, restricted to the three closely-related species, and the inferred distribution of Stalker2 from the past, in which it was present in five species, are both consistent with vertical transmission. However, Stalker2 is a striking candidate for horizontal transfer, with incredibly small divergence between elements for all three interspecies comparisons, ranging from as little as 0.001 between *D. simulans* and both *D. melanogaster* and *D. sechellia*, to 0.002 between *D. melanogaster* and *D. sechellia*. However, despite these striking results, the maximum parsimony tree produced is congruent with the host phylogeny, with all elements from *D. melanogaster* forming one clade, with all elements from *D. simulans* and *D. sechellia* forming a sister clade. The maximum likelihood tree is incongruent, with a single element from *D. sechellia* found within the *D. melanogaster* clade. The inconsistency between these two trees is due to the incredibly small divergence between sequences, and therefore the lack of mutational data to infer relationships. This renders the trees unreliable and therefore uninformative, but the strength of divergence data alone strongly support the hypothesis of horizontal transfer, most likely incredibly recent and involving all three species, although the direction of transfer cannot be inferred. Divergence is too small to be explained by transfers involving the *D. simulans*/*D. sechellia* ancestor, and therefore at least three recent transfers can be inferred. The phylogenies produced using both flanking and solo LTR sequences from *D. melanogaster*, *D. simulans* and *D. sechellia*, and the solo LTR sequences from *D. yakuba* and *D. erecta*, are congruent with the known host relationships, and therefore do not suggest that horizontal transfer may have occurred involving *D. yakuba* or *D. erecta* prior to the elimination of

Stalker2 from these species. Closer examination of the sequences of the Stalker2 LTR provides further evidence to support this. A deletion between positions 233 and 244 which is found in ten out of the thirteen solo LTR sequences in *D. yakuba* is not found in the LTRs of any other species, providing evidence against transfer from *D. yakuba*. There is a 6bp deletion towards the 3' end of twelve of the sixteen *D. erecta* LTRs, which again, is unique to this species. Divergence data also suggest horizontal transfer involving these two species has not occurred, with more than 10% divergence between LTRs in these species and those in their close relatives *D. melanogaster*, *D. simulans* and *D. sechellia*.

Tabor is a Gypsy-like family, which is an interesting family to investigate as like Stalker2, in addition to full-length elements found in *D. melanogaster*, *D. simulans* and *D. sechellia*, solo LTRs are found in *D. yakuba*. A distantly-related family, Tabor_DA, is found in *D. ananassae*, but elements, and their LTRs, are completely absent from *D. erecta*. If Tabor_DA is indeed descended from the same ancestral element as Tabor, this would suggest vertical transmission of the element, followed by stochastic loss of both complete elements and solo LTRs along the *D. erecta* lineage, and loss of complete elements, and any evidence of the internal portion, along the *D. yakuba* lineage. However, the current distribution of Tabor, which is restricted to *D. melanogaster*, *D. simulans* and *D. sechellia*, is not patchy, and does not support the hypothesis of horizontal transfer. The phylogenies produced are congruent with the host phylogeny. However, this does not preclude the possibility of horizontal transfer. It is possible that, as in the case of *D. yakuba*, elements have been entirely lost from particular species and have been reintroduced by horizontal transfer. In this instance, the internal trees would match the host phylogeny. Divergence data strongly suggest that Tabor has been involved in at least two recent horizontal transfer events, with smallest divergence of 0.002 for all three interspecies comparisons. Due to the presence of solo LTRs in *D. yakuba*, the relationships between Tabor LTRs were investigated further. The case is a lot more complicated than for Stalker2, as divergence between the LTRs in *D. yakuba* and those from the other species are all smaller than for the host genes. The phylogeny produced using the sequences of the flanking and solo LTRs of *D. melanogaster*, *D. simulans* and *D. sechellia*, and the solo LTRs from *D. yakuba*, is, however, congruent with the host tree. All LTRs from *D. yakuba* cluster together to the exclusion of LTRs from the other species, however, this is to be expected due to the extremely close relationships between elements of these three species, discussed above. To investigate the possibility of horizontal transfer further, the Tabor LTR sequences were examined in detail. There is a 15bp deletion ("deletion 1") linked to a point mutation, suggesting it has only occurred once, found in two out of the seventeen *D. yakuba* LTRs, one of the seven *D. simulans* LTRs, no *D. melanogaster* LTRs and all *D. sechellia* LTRs. This may represent an ancestral polymorphism that has been retained. A second interesting deletion is 11bp long, and found in all *D. yakuba* LTRs ("deletion 2"). The fact that two *D. yakuba* LTRs possess both deletions could be explained by the *D. yakuba*-specific deletion occurring multiple times in parallel, which is possible as the sequence that has been deleted is highly similar to the sequence directly upstream from it. However, if this were the case it might be expected that the mutation would occur in other species. Alternatively, recombination may have occurred. Although the pattern of the first deletion suggests ancestral polymorphism, the level of divergence in the internal part of the element is probably too low for this to be a possibility, at least in the case of flanking LTRs. Assuming the first deletion only happened once, as it, and the linked point mutation, are found in the LTRs of *D. yakuba*, *D. simulans* and *D. sechellia*, it appears that this deletion occurred before these species diverged, around eight million years ago. Therefore the elements in *D. sechellia*, which all have this deletion in their flanking LTRs, can perhaps be expected to be at least eight million years divergent from the Tabor elements of *D. melanogaster*, the LTRs of which do not contain this deletion. However, in the most extreme case, the amount of divergence between a *D. melanogaster* element and a *D. sechellia* element is less than 1%. Since the LTRs can recombine, it is possible that the internal portions of these elements are very closely related, which is why they have such a high degree of sequence conservation, and that the LTRs from the other type have been brought in by recombination. If the deletions do indeed represent ancestral polymorphism, there are three possibilities. The first explanation is ancestral polymorphism of deletion 1, which has been lost in *D. melanogaster* and fixed in *D. sechellia*. Both forms have been retained in *D. simulans* and recombination, followed by loss of one type, has occurred in *D. yakuba*. The second is ancestral polymorphism of deletion 2, which would again require either recombination to have occurred in *D. yakuba*, or multiple occurrences of deletion 2. The third is ancestral polymorphism whereby two forms exist, a full-length LTR and one containing both deletion 1 and deletion 2. Recombination would need to have occurred to generate the variety of sequences observed, however, this possibility would require more steps and is a less parsimonious explanation for the observation. The most likely explanation for the variation among Tabor LTRs appears to be ancestral polymorphism of deletion 1, followed by multiple occurrence of deletion 2 due to replication slippage. It is impossible to determine the ancestral state of the Tabor LTR, as the alignment with the most closely-related element, Tabor_DA, is poor and uninformative. Horizontal transfer involving *D. yakuba* is not supported, due to the presence of deletion 2, which is found in all *D. yakuba* LTRs, and does not occur in any of the other species.

The extent of the divergence between elements of both the **Frogger** and **Gypsy9** families does not suggest horizontal transfer of these families has occurred within *D. melanogaster*, *D. simulans* and *D. sechellia*.

4.3.4.3 LTR retrotransposons with more extensive distribution throughout the Sophophora, but absent from the Drosophila subgenus

There are nine families of LTR retrotransposons distributed throughout the Sophophora but absent from the Drosophila subgenus for which horizontal transfer is supported by three pieces of evidence: small divergence between elements of the same family in different species, phylogenetic incongruence, and a patchy distribution across the host phylogeny. These families are Batumi, Bel, Blastopia, Circe, Copia, Diver, Diver2, Gypsy5 and Max.

Batumi is restricted to *D. melanogaster*, *D. simulans*, *D. sechellia* and *D. yakuba*. Batumi is unexpectedly absent from *D. erecta*. Small divergence supports horizontal transfer between *D. yakuba* and all of the other species, and also between *D. simulans* and *D. melanogaster*. However, no phylogenetic incongruence involving *D. yakuba* elements is observed, as these elements cluster together to the exclusion of all others. A group of elements from *D. melanogaster* falls within the *D. simulans*/*D. sechellia* clade, perhaps suggesting transfer from the ancestor of *D. simulans* and *D. sechellia* into *D. melanogaster*. Precise relationships, however, are inconsistent between the two phylogenetic construction methods. The small divergence observed between Batumi elements when comparing *D. melanogaster*, *D. simulans* and *D. sechellia* with *D. yakuba* could be explained by a single transfer event involving *D. yakuba* and the ancestor of the other three species.

The **Bel** family has the same distribution as Batumi. Divergence between elements is strikingly small, with all interspecies comparisons, with the exception of *D. simulans*/*D. sechellia*, yielding smallest divergence less than that of *Adh*. Although the range of divergence can be up to as much as 18% (*D. sechellia*/*D. yakuba*), the majority of comparisons indicate little divergence. The observation of small divergence for so many interspecies comparisons may indicate multiple transfer events, as the values are so small that they indicate a number of very recent transfers, rather than a single transfer resulting in high identities due to relationships between species. Such a scenario is supported by the phylogenies produced. The closest relatives of the *D. yakuba* elements in the tree are those from *D. melanogaster*. The direction of transfer is ambiguous, as the elements simply form a clade together, however, transfer from *D. melanogaster* into *D. yakuba* may be more likely as this would account for the high identity between elements in *D. yakuba* compared with *D. simulans* and *D. sechellia*. An additional transfer appears to have occurred from *D. simulans* into *D. melanogaster*, as elements from the latter species are found within the clade of elements from the former. Transfer in this direction would explain the high identity between elements in *D. melanogaster* and *D. sechellia*, as *D. simulans* and *D. sechellia* are very closely-related species. The divergence values support these two transfers, with the smallest divergence observed between *D. simulans* and *D. melanogaster* 0.003, and between *D. melanogaster* and *D. yakuba* 0.006. Divergence between *D. melanogaster* and *D. sechellia* is slightly higher at

0.008, and divergence values between *D. simulans* and *D. sechellia* compared with *D. yakuba* are also higher, at 0.013 and 0.011, respectively.

Blastopia is present in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. ananassae*, *D. pseudoobscura* and *D. persimilis*. This distribution is patchy as the family is absent from the other members of the melanogaster group, *D. yakuba* and *D. erecta*. Smallest divergence obtained for each of the interspecies comparisons of the internal section of the Blastopia sequence within the *melanogaster* subgroup are considerably lower than the corresponding values for *Adh* and are likely to be explained by horizontal transfer among the species of the *melanogaster* subgroup. This is further supported by phylogenetic incongruence, with incongruent relationships between *D. melanogaster*, *D. simulans* and *D. sechellia* internal sequences consistently supported by both phylogenetic construction methods. The direction of any transfer events cannot be elucidated from the internal phylogenies, however, the phylogenies produced using LTR sequences support the hypothesis of transfer from *D. melanogaster* to *D. simulans*, with *D. simulans* elements nested within the *D. melanogaster* clade. There is no evidence from either the extent of divergence between elements, or the phylogenies constructed, to support horizontal transfer into the *melanogaster* subgroup as an explanation for the absence from *D. yakuba* and *D. erecta*. Therefore, it is possible that the patchy distribution of this family can be attributed to stochastic loss.

The distribution of **Circe** is inconsistent with vertical transmission, found in the close relatives *D. melanogaster*, *D. simulans*, *D. sechellia* and *D. yakuba,* but absent from *D. erecta*. Small divergence between elements in *D. melanogaster* and *D. yakuba*, which can be as small as 0.036 (compared with

0.044 for *Adh*), suggests horizontal transfer may have occurred involving these two species. This is supported by the phylogenies of Circe elements, in which elements from *D. yakuba* are consistently found within a cluster of *D. melanogaster* elements, indicating horizontal transfer involving *D. melanogaster* as the donor and *D. yakuba* as the recipient. This would explain the absence of this family from *D. erecta*.

It has been suggested previously that the **Copia** family may have undergone horizontal transfer (Sanchez-Gracia et al. 2005). Copia has an unusual distribution, found in *D. melanogaster*, *D. simulans* and *D. sechellia* of the *melanogaster* subgroup, in addition to the distant relative *D. willistoni*. Copia has proliferated to a great extent in *D. melanogaster*, with a high copy number, and the majority of copies identical to each other, suggesting a rapid, and recent, proliferation. In contrast, no full-length copies of Copia were extracted from the *D. sechellia* genome. Although there is considerable divergence between Copia elements in *D. willistoni* and those found in the other species, this divergence is smaller than that of the *Adh* gene. Horizontal transfer of Copia is also supported by the neighbour-joining phylogeny produced by Sanchez-Gracia et al. (2005). However, a neighbour joining phylogeny produced using the elements extracted in this study from the complete genome sequences of *D. melanogaster*, *D. simulans* and *D. sechellia* clustered all *D. simulans* and *D. sechellia* elements together to the exclusion of *D. melanogaster* elements with bootstrap support of 100%. The maximum likelihood and maximum parsimony phylogenies also support this relationship, which is congruent with the host phylogeny, contrary to the findings of Sanchez-Gracia et al. The relationship is supported in 97.4% of bootstrap replicates, and further supported by examination of shared mutations in the sequences. There are fourteen variable sites which group all

D. simulans and *D. sechellia* elements together to the exclusion of elements from *D. melanogaster*. Therefore, although small divergence is observed between elements belonging to these three species, the phylogenies produced in this analysis do not provide additional support for the hypothesis of horizontal transfer. Phylogenetic incongruence is observed, however, when the Copia sequence from *D. willistoni* is included (Figure 4.15). In both the maximum likelihood and maximum parsimony trees, the *D. willistoni* element groups with *D. simulans* and *D. sechellia* to the exclusion of *D. melanogaster*, with high levels of support (0.92 and 94% respectively). Copia LTR sequences have not diverged to a lesser extent than expected under vertical transmission.

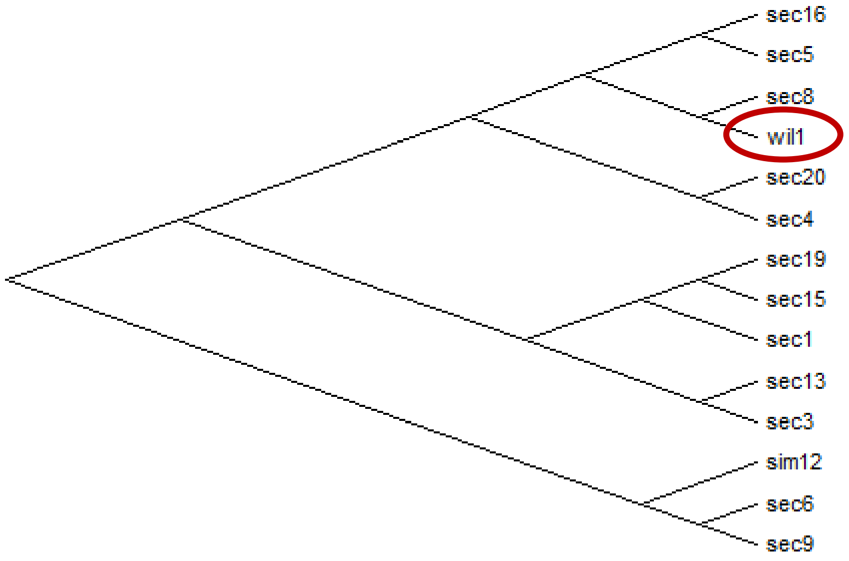


Figure 4.15: Section of the Copia maximum parsimony phylogeny, showing the incongruent position of the element from *D. willistoni* within a clade of elements from *D. simulans* and *D. sechellia*.

Similarly to Bel, the divergence observed between elements of the **Diver** family in different species suggests multiple, very recent, horizontal transfer events may have occurred. Diver is found in the close relatives *D.*

melanogaster, *D. simulans*, *D. sechellia* and *D. yakuba*, and is therefore unexpectedly absent from *D. erecta*. Every interspecies comparison of Diver elements yields a smallest p distance less than that of the host gene *Adh*. The divergence in each case is strikingly small, from 0.001 between *D. melanogaster* and *D. yakuba*, to 0.008 between *D. melanogaster* and *D. simulans*. A limited number of transfers combined with relationships between species, except potentially the very close relatives *D. simulans* and *D. sechellia*, are therefore unlikely to explain these values, and multiple transfers can be inferred to have occurred. However, due to the extremely small amount of mutation data available for phylogenetic reconstruction, the two methods employed to produce phylogenies generate inconsistent results. Both the maximum parsimony and maximum likelihood trees are incongruent with the known host phylogeny, however, specific relationships are inconsistent between the two trees. On the parsimony tree, the position of elements from *D. yakuba* is congruent with the host phylogeny, whereas on the likelihood tree these elements are found among elements from other species. Incongruence involving *D. melanogaster*, *D. simulans* and *D. sechellia* is consistently observed on both phylogenies. The relationships observed cannot be considered reliable, and therefore the trees are uninformative in deducing the likely direction of horizontal transfer events. However, the strikingly small divergence between elements in different species provides convincing evidence for horizontal transfer in the absence of supporting evidence from the phylogenies.

Diver2 is found in the same species as Diver. It is not as striking a candidate for horizontal transfer, with divergence smaller than the host genes for only two interspecies comparisons, which in both cases is not indicative of very recent transfer. Smallest divergence observed between *D. melanogaster* and

D. yakuba is 0.042, compared with 0.044 for *Adh*, and between *D. simulans* and *D. yakuba* is 0.037, compared with 0.038 for *Adh*. Divergence between *D. sechellia* and *D. yakuba* is also 0.037, although in this case exceeds the value for *Adh*, 0.035, but is clearly worthy of consideration. This case illustrates one of the problems associated with comparison with host genes as a means of inferring horizontal transfer, as this is clearly not an appropriate "cut-off" to confidently infer transfer either has, or has not, occurred. The divergence time between *D. yakuba* and both *D. simulans* and *D. sechellia* is identical, as is the smallest divergence between Diver2 elements, however, were this cut-off to be adhered to, only transfer between *D. simulans* and *D. yakuba* would be inferred. Two clusters containing elements from *D. melanogaster* and *D. yakuba* are consistently observed in two different positions on the phylogenies produced. One cluster is particularly well-supported, in 89% of bootstrap replicates on the maximum parsimony tree. The direction of transfer appears to be from *D. melanogaster* to *D. yakuba*, as the elements from *D. yakuba* are found within a cluster of elements from *D. melanogaster*. Such a transfer would also explain the low divergence observed between *D. simulans* and *D. yakuba*, as an artefact of the close relationship between *D. simulans* and *D. melanogaster*. The phylogenies also demonstrate incongruent relationships involving *D. melanogaster* and *D. sechellia*, with a possible transfer inferred from *D. melanogaster* into *D. sechellia*. However, such a transfer is not supported by the divergence between elements, and therefore, if it did indeed occur, is likely to be relatively ancient. An alternative explanation for this incongruence could be differential retention of an ancestral polymorphism.

Gypsy5 is a low copy number member of the Gypsy superfamily, present in *D. melanogaster*, *D. simulans*, *D. yakuba* and *D. erecta*. Absence of this family from *D. sechellia* is unexpected, given it is found in the close relative *D. simulans*. However, due to the low copy number of the family, which has only one element in both *D. simulans* and *D. melanogaster*, it is perhaps more likely that the patchy distribution of this family can be explained by stochastic loss from *D. sechellia* rather than horizontal transfer. However, small divergence between elements for four separate interspecies comparisons: *D. melanogaster*/*D. simulans* (0.017), *D. melanogaster*/*D. yakuba* (0.004), *D. simulans*/*D. yakuba* (0.004) and *D. yakuba/D.erecta* (0.021), strongly supports horizontal transfer of this family. The phylogenies of the internal sequences consistently add weight to a potential transfer from *D. yakuba* to *D. erecta*, with the single *D. erecta* element nested in the *D. yakuba* clade. There are insufficient copies of the family in the other two species to provide further evidence for horizontal transfer from the phylogeny (Figure 4.16), or to deduce a likely direction of any such transfer. Copy number is higher for solo LTRs, but small divergence and short sequence length makes relationships difficult to resolve. Although both maximum likelihood and maximum parsimony trees of the LTRs provide numerous examples of incongruence, the two trees are inconsistent and therefore cannot be reliably interpreted.

Max has a relatively limited distribution, restricted to *D. melanogaster*, *D. simulans*, *D. sechellia* and *D. yakuba*, and is unexpectedly absent from *D. erecta*. Smallest divergence between elements of the Max family in different species is less than that between the host *Adh* genes for all interspecies comparisons. These values range from as small as 0.004 for the *D. melanogaster* comparison with *D. simulans*, to 0.014 for the comparison of *D. melanogaster* and *D. yakuba*, providing strong evidence in support of horizontal transfer of the Max family. Transfer from *D. simulans* into *D. melanogaster* is supported as four elements from *D. melanogaster* cluster with an element from *D. simulans* within the clade of all elements from *D. simulans* and *D. sechellia*. Given that *D. simulans* and *D. sechellia* are extremely closely related, a transfer in this direction would also yield small divergence between elements in *D. melanogaster* and *D. sechellia*, in the absence of direct transfer between them, which is indeed observed. Transfer between *D. simulans* and *D. sechellia* cannot be confidently supported by the phylogenies, due to the short divergence time separating these two species. No phylogenetic incongruence is observed involving *D. yakuba*, as elements from this species all cluster together to the exclusion of all other elements. However, this does not eliminate the possibility of horizontal transfer involving this species, as such a transfer may have introduced, or reintroduced, the family to this species, which would result in all elements clustering together.

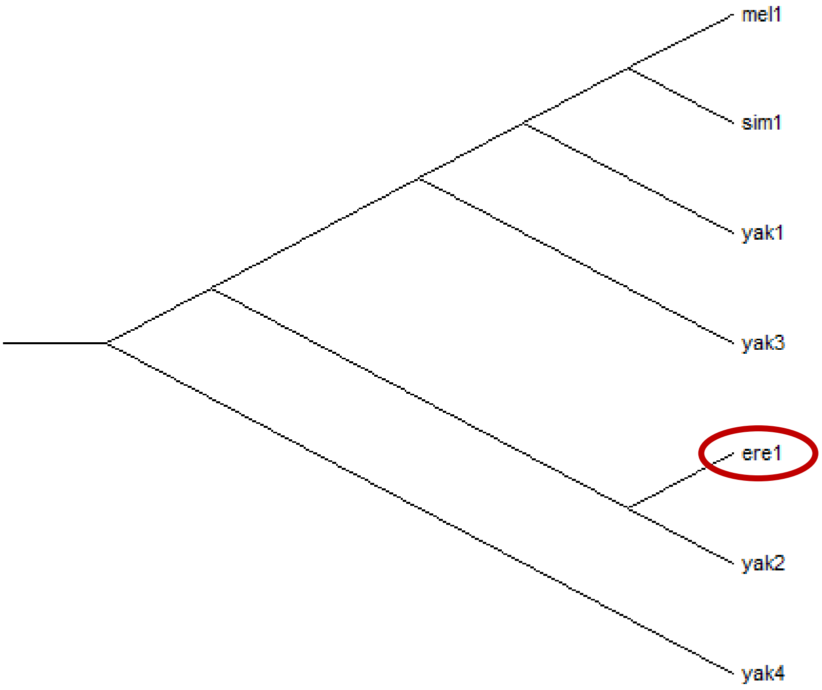


Figure 4.16: Maximum parsimony phylogeny of Gypsy5 elements, showing phylogenetic incongruence.

There are eighteen families of LTR retrotransposons distributed throughout the Sophophora but absent from the Drosophila subgenus for which horizontal transfer is supported by two of the three pieces of evidence described above.

All cases are supported by small divergence, with additional support due to either phylogenetic incongruence or patchy distribution.

The distribution of the **297** family is consistent with vertical transmission. It has an extensive distribution in the Sophophora, found in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta* and *D. ananassae*. Despite its relatively wide distribution, divergence between elements is only smaller than for *Adh* when comparing *D. melanogaster* with its closest relatives *D. simulans* and *D. sechellia*, with the smallest divergence observed 0.004 and 0.011, respectively. Transfer between *D. melanogaster* and *D. simulans* is supported by the constructed phylogenies, as a *D. simulans* element consistently clusters with a group of three *D. melanogaster* elements, within the entire *D. melanogaster* clade. This suggests transfer occurred from *D. melanogaster* into *D. simulans*. Transfer in this direction, however, does not explain the high identity shared between *D. melanogaster* and *D. sechellia*. Were the transfer to have occurred from *D. simulans* into *D. melanogaster*, similarity with *D. sechellia* would be expected based on the close relationship between *D. simulans* and *D. sechellia*, and therefore high identity expected between elements of these species. This observation is therefore explained by a second transfer event involving *D. sechellia*, which is supported by the phylogeny by the clustering of an element from *D. sechellia* just outside the *D. melanogaster* clade, rather than with the other elements from *D. simulans* and *D. sechellia*. Again, this suggests a direction of transfer involving *D. melanogaster* as the donor species. Transfer of 297 between *D. melanogaster* and both *D. simulans* and *D. sechellia* has been proposed (Vidal et al. 2009), but the direction of transfer was not inferred.

Burdock has a more extensive distribution, found in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta* and *D. ananassae*, again consistent with vertical transmission, as these species are each other's closest relatives. Five interspecies comparisons yield smallest divergence less than that for the corresponding *Adh* genes, which are the comparisons of *D. melanogaster* and *D. yakuba* with both each other, and *D. simulans* and *D. sechellia*. The most striking is the comparison of *D. melanogaster* and *D. simulans*, for which the smallest divergence observed is only 0.004. Transfer involving these two species is supported by the phylogenies produced, in which a single *D. simulans* element is found within a group of closely-related *D. melanogaster* elements, suggesting transfer from *D. melanogaster* into *D. simulans*. Additionally, a group of elements from *D. simulans* and *D. sechellia* is found within the *D. melanogaster* clade as a whole. This may indicate transfer from *D. melanogaster* into the ancestor of *D. simulans* and *D. sechellia*, or into each of these species individually. The tree therefore supports two of the hypothetical transfers suggested by the divergence observed between elements in different species. However, no incongruence is observed involving *D. yakuba*. Elements from this species cluster together in the appropriate place on the tree. It is possible that a transfer may have occurred into the *D. yakuba* lineage, following stochastic loss of the family from this species, prior to the diversification of *D. melanogaster*, *D. simulans* and *D. sechellia*. The divergence observed between elements is consistent with this hypothesis, as are the relationships observed in the phylogenies. It seems unlikely that vertical transmission, followed by such small divergence occurring between all three combinations of species, would occur.

The **Gypsy** family is a famous known example of horizontal transfer (reviewed by Loreto et al. 2008). Gypsy was also included in this investigation of horizontal transfer, with the advantage of access to the complete genome sequences for twelve *Drosophila* species. Gypsy was identified in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*, but not in any of their more distant relatives. This distribution is consistent with vertical transmission. However, as anticipated, evidence for horizontal transfer of Gypsy was found through both divergence data and phylogenetic incongruence. Divergence between Gypsy elements in different species is smaller than that between the host *Adh* genes for the three interspecies comparisons involving only *D. melanogaster*, *D. yakuba* and *D. erecta*. No evidence was found to suggest horizontal transfer involving *D. simulans* or *D. sechellia*. Between *D. yakuba* and *D. erecta*, divergence was as little as 0.018, and the phylogenies produced support transfer between these species, as *D. yakuba* elements are found within the *D. erecta* clade, supporting transfer from *D. erecta* into *D. yakuba*. *D. melanogaster* elements are also found within the *D. erecta* clade, again supporting transfer involving *D. erecta* as the donor species. The high degree of similarity between *D. melanogaster* and *D. yakuba*, which can be as high as 98%, may be attributed to the two transfers from *D. erecta* into both of these species. Alternatively, further transfer may have occurred between *D. melanogaster* and *D. yakuba*, the elements from which do cluster together within the *D. erecta* clade.

Gypsy2 is found in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*, and therefore does not follow a patchy distribution across the *Drosophila* phylogeny. In the case of Gypsy2, only a single interspecies comparison yields divergence smaller than that of the host gene *Adh*, that is the comparison between *D. melanogaster* and *D. erecta* (0.026). All elements from *D. melanogaster* do indeed fall within the *D. erecta* clade on the phylogenies, perhaps suggesting Gypsy2 was introduced, or reintroduced having been stochastically lost, to *D. melanogaster* from *D. erecta*, with all elements now present descended from the transfer event. Interestingly, within the *D. erecta* clade, the elements from *D. yakuba* are also found, and are most closely related to the elements from *D. melanogaster*. Transfer from *D. melanogaster* to *D. yakuba* is indeed plausible, with the smallest divergence observed only slightly in excess of the equivalent value for *Adh* (0.049, compared with 0.044). To explain these relationships, transfer from *D. erecta* to *D. melanogaster* would have to occur first, followed by transfer from *D. melanogaster* to *D. yakuba*.

Gypsy4 is found in all of the species in which Gypsy2 is found, along with *D. ananassae*, consistent with vertical transmission. However, in two cases, divergence observed is smaller than the host genes, indicating possible horizontal transfer. The smallest divergence between elements in *D. melanogaster* and *D. simulans* is 0.014, and between *D. yakuba* and *D. erecta* is 0.007, considerably lower than would be expected given the divergence time between these two species. Horizontal transfer as an explanation for this small divergence is supported by the phylogenies. All elements from *D. yakuba* form a clade with all elements from *D. erecta*, with a score of 1.0 on the maximum likelihood tree, however the direction of transfer cannot be inferred. It appears that this transfer reintroduced the family into a species from which it had been lost, or that more ancient elements, not descended from the transfer, have been lost in the time since the transfer event, as a separate clade for each species is not observed. Loss of the family from one of these two species, rather than the family simply having always been absent, is supported by the presence of Gypsy4 in *D. ananassae*, for which there is no evidence of horizontal transfer. All elements from *D. simulans* and *D. sechellia* are found within the *D. melanogaster* clade, thereby suggesting transfer from *D. melanogaster* introduced the family into the ancestor of these two species. Once again, the relationships in the tree are indicative of loss of the family followed by reintroduction by horizontal transfer.

Gypsy10 is found in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*, and, despite a distribution reflecting vertical transmission, is very likely to have been involved in recent horizontal transfer. The smallest divergence between elements in different species is less than that between the host *Adh* genes in all cases except for all comparisons involving *D. simulans*. The phylogenies produced are incongruent with the host tree and further support the hypothesis of horizontal transfer. *D. melanogaster* elements fall within a clade of *D. erecta* elements, suggesting transfer between these species with *D. erecta* as the donor. Surprisingly, two elements from *D. simulans* are also found within the *D. yakuba*/*D. erecta* clade.

The least convincing case for horizontal transfer among the LTR retrotransposons comes from the **Gypsy12** family. This family is present in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. ananassae*. Absence of Gypsy12 from *D. erecta* results in a patchy distribution across the phylogeny, and renders Gypsy12 a potential family to have been involved in horizontal transfer. This may have occurred between one of the four closest relatives out of these species (*D. melanogaster*, *D. simulans*, *D. sechellia* and *D. yakuba*), or their common ancestor, and their more distant relative *D. ananassae*. However, divergence data do not support this theory. Divergence for no interspecies comparisons is smaller than *Adh*, although divergence is equal to that of *Adh* for the *D. melanogaster* comparison with *D. simulans*, which may suggest that transfer has occurred between these two species. However, the phylogenies produced are congruent with known host relationships, providing no further support for transfer between *D. melanogaster* and *D. simulans*, nor transfer involving *D. ananassae*.

Invader1 has a relatively extensive distribution, found in six of the nine species of the Sophophora for which the sequenced genomes are available: *D. melanogaster* and its closest relatives, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta* and *D. ananassae*. This distribution is consistent with vertical transmission, but horizontal transfer does appear to have occurred. Divergence is smaller than that of the host gene *Adh* when comparing elements from *D. melanogaster* and *D. simulans* (0.008), and *D. yakuba* with *D. erecta* (0.049). Elements from these species fall in various parts of the tree, suggesting that any transfer that occurred did not introduce the family into a naïve genome. This is also supported by the distribution of the family. Incongruent relationships supporting both transfers suggested by divergence data are consistent on both phylogenies. Elements from *D. yakuba* and *D. erecta* cluster together in three separate locations on the tree. One of these is nestled within the *D. melanogaster* clade, perhaps indicating transfer from *D. melanogaster* into one of these species. Divergence between *D. melanogaster* and both *D. yakuba* and *D. erecta* is equally small. Additionally, a single element from *D. ananassae* is found in an incongruent position on the tree, which may reflect a relatively ancient transfer not supported by divergence data.

The distribution of **Invader4** is consistent with vertical transmission, with the family found in the genomes of *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*. However, divergence between elements suggests that horizontal transfer is likely to have occurred between *D. yakuba* and *D. erecta*, as there are two pairs of elements in the species which are 100% identical to each other. This is extremely unlikely to occur by chance given the divergence time between these two species, and the low copy number of the family in both species. The phylogenies produced consistently support the hypothesis of transfer between *D. yakuba* and *D. erecta*. A mixed clade of *D. yakuba* and *D. erecta* elements is found on the tree, with the clade, and all individual relationships within it, supported in 100% of bootstrap replicates on the maximum parsimony tree (Figure 4.17). It is not possible to infer the direction of transfer from the phylogenies.

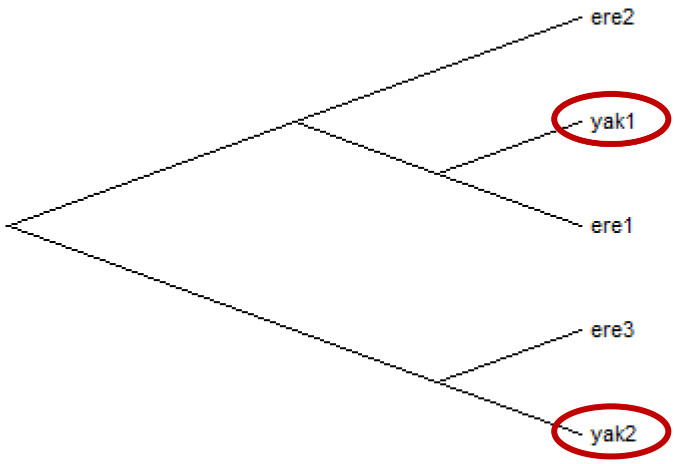


Figure 4.17: Section of the Invader4 maximum parsimony phylogeny, showing a mixed clade of elements from *D. yakuba* and *D. erecta*.

Mdg1 is another family which is distributed in a manner consistent with vertical transmission, present in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*. Horizontal transfer of Mdg1 is, however, supported by both small divergence between elements in different species, and by incongruence of the Mdg1 phylogenies with known host relationships. Evidence for horizontal transfer from divergence data is not particularly striking for Mdg1, with only two interspecies comparisons yielding divergence smaller than *Adh*, and not substantially so. Smallest divergence between *D. melanogaster* and *D. sechellia* elements is 0.014, compared with 0.017 for *Adh*, and between *D. melanogaster* and *D. yakuba* is 0.032, compared with 0.044 for *Adh*. However, the latter of these putative transfers is not supported by the phylogenies produced. Surprisingly, two elements from *D. erecta* are found within the cluster of *D. yakuba* elements, perhaps suggesting transfer from *D. yakuba* into *D. erecta* has occurred. Smallest divergence between Mdg1 elements in these species is small, but not smaller than between the *Adh* genes (0.090 compared with 0.051). Transfer between *D. melanogaster* and *D. sechellia*, which was indicated by the divergence data, is supported by the phylogenies produced. The majority of elements from *D. sechellia* fall within the *D. melanogaster* clade, supporting transfer involving *D. sechellia* as the recipient species. Transfer in this direction is consistent with divergence greater than *Adh* between *D. melanogaster* and *D. simulans*.

Mdg3 follows a relatively restricted distribution, found in *D. melanogaster*, *D. simulans*, *D. sechellia* and *D. yakuba*, once again inconsistent with vertical transmission from a common ancestor due to the absence of the family from *D. erecta*. Divergence between elements suggests that horizontal transfer may have occurred, but there is no further evidence to support this hypothesis as the phylogenies produced are congruent with the known host relationships. Divergence between elements can, however, be strikingly small for the Mdg3 family, ranging from as little as 0.009 between *D. melanogaster* and *D. yakuba*. Three other interspecies comparisons, comparing *D. melanogaster* and *D. simulans*, *D. simulans* and *D. yakuba*, and *D. sechellia* and *D. yakuba* also yield smallest divergence less than for *Adh*. It is possible that recent transfer has occurred from *D. melanogaster* into *D. yakuba*, as this would result in small divergence being observed between *D. yakuba* and the other two species, due to their close relationship with *D. melanogaster*, and would explain the absence of Mdg3 from *D. erecta*. Mdg3 has been shown to be capable of horizontal transfer *in vitro* (Syomin et al. 2002).

The **Nomad** family has an interesting evolutionary history, whereby, like Hobo of the DNA transposons, many sequences appear to have propagated non-autonomously. In these cases, only the 5' and 3' ends of the elements are intact, with the middle portion of the element having been deleted. Presumably, the sequences required for recognition of Nomad by its enzymes are found in these retained regions, and enzymes produced by full-length elements can be inferred to have mobilised the shortened form of the element. The distribution of Nomad elements is consistent with vertical transmission, with the family found in the genomes of *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta* and *D. ananassae*. Despite this relatively wide distribution, divergence between elements is also generally consistent with vertical transmission, and is only smaller than for *Adh* when comparing *D. melanogaster* with its two closest relatives, *D. simulans* and *D. sechellia*. The smallest divergence observed in both cases is very small, at only 0.007. The phylogenies produced support horizontal transfer involving these species. The entire clade of *D. melanogaster* elements falls within the *D. simulans*/*D. sechellia* clade. This observation, and the divergence observed between elements, supports the hypothesis of horizontal transfer from the ancestor of *D. simulans* and *D. sechellia* into *D. melanogaster*.

Quasimodo has an unusual distribution, found in only two species which are not each other's closest relatives, *D. melanogaster* and *D. erecta*. For this patchy distribution to be observed under the null hypothesis of vertical transmission, the family must have been present in the ancestor of these species and then stochastically lost a least twice: from the *D. yakuba* lineage and from the *D. simulans*/*D. sechellia* lineage. Alternatively, the distribution could be explained by horizontal transfer introducing the family into one of the two species from the other. This is supported by the divergence observed between elements in the two species, which can be as small as 0.028, compared with 0.049 for *Adh*. The phylogenies produced are congruent with the host phylogeny in that all elements from *D. erecta* form a clade to the exclusion of all elements from *D. melanogaster*. However, this relationship would also be expected if horizontal transfer had introduced this family into a naïve genome, as all elements within the genome would be descended from the transfer event, and would form a clade together in the tree. This may be expected to fall within the clade of elements from the other species, however, as Quasimodo is only found in two species, whether or not this is the case would depend on the position of the root, and therefore cannot be inferred. Given the evidence, it is likely that horizontal transfer of Quasimodo has occurred between *D. melanogaster* and *D. erecta*, however, the direction cannot be inferred. The absence of Quasimodo from *D. simulans* and *D. sechellia*, which are closely related to *D. melanogaster*, suggests transfer from *D. erecta* into *D. melanogaster* as a more likely route, however, absence from *D. yakuba* supports transfer in the opposite direction. The family could have been stochastically lost from the *D. simulans*/*D. sechellia* lineage, or may have been introduced into *D. melanogaster* from an unknown donor. Stochastic loss of Quasimodo from *D. simulans* and *D. sechellia*, or the ancestor of these two species, is confirmed by the presence of solo LTRs in both of these species. Solo LTRs are absent from *D. yakuba*, however, therefore it is not clear whether or not absence from this species can be attributed to stochastic loss or to horizontal transfer. The presence of solo LTRs in *D. simulans* and *D. sechellia* counteracts the argument that absence of Quasimodo from these species supports horizontal transfer involving *D. melanogaster* as the recipient species. The solo LTRs in *D. simulans* and *D. sechellia* have diverged considerably from those of *D. erecta*, however, this does not provide significant evidence relating to the direction of transfer, but does suggest that transfer involving *D. simulans*, *D. sechellia* and *D. erecta* did not occur prior to the loss of Quasimodo from all but the latter of those species. Examination of the Quasimodo LTR sequences reveals no mutations which suggest horizontal transfer between *D. simulans* or *D. sechellia* and *D. erecta*. The phylogenies constructed using Quasimodo LTR sequences are as expected, with LTRs from *D. melanogaster* forming a clade with elements from *D. erecta*.

Roo has an extensive distribution, present in eight out of the nine species from the Sophophora subgenus for which the complete genome sequence is available, absent only from *D. willistoni*. However, Roo appears to be in the process of being eliminated from the *D. persimilis* genome (de la Chaux and Wagner 2009). The family is also absent from the Drosophila subgenus. Divergence between elements in different species is smaller than for the host gene *Adh* for the equivalent comparison for nine interspecies pairs, involving only the closely-related species *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*. Elements of the Roo family in *D. ananassae*, *D. pseudoobscura* and *D. persimilis* do not appear to have been involved in recent horizontal transfer events. Smallest divergence ranges from 0.000, whereby elements in the two species are identical to each other in *D. melanogaster* and *D. sechellia*, to 0.043 between *D. melanogaster* and *D. erecta*, which is not substantially smaller than the divergence in *Adh*, 0.049. There are several examples of consistent incongruent relationships involving these five species on the phylogenies constructed. *D. pseudoobscura*, *D.*

persimilis and *D. ananassae* elements fall in positions which are congruent with the host relationships. Transfer appears to have occurred from *D. melanogaster* into *D. sechellia*, as an element from *D. sechellia* is found in a well-supported position in the *D. melanogaster* clade. The elements of the two species that cluster together share 100% identity. All elements from *D. erecta* fall into a clade of elements from *D. yakuba*, perhaps suggesting the Roo family had been lost from *D. erecta* and was repopulated by transfer from *D. yakuba*. Alternatively, it could simply be that all elements present in the *D. erecta* genome sequence happen to be descended from the transfer event. On the maximum parsimony tree, an element from *D. simulans* falls within the *D. melanogaster* clade, but this relationship is not supported on the maximum likelihood tree. However, elements from *D. simulans* do cluster more closely with those from *D. melanogaster* than those from *D. sechellia* on both trees, supporting the hypothesis of transfer between these species. Transfer between *D. sechellia* and *D. yakuba* is also supported, with elements from *D. yakuba* found within the *D. sechellia* clade, which indicates *D. sechellia* as the likely donor species. Interestingly, the clade of *D. yakuba*/*D. erecta* elements is more closely related to the *D. melanogaster*/*D. simulans* clade than is the *D. sechellia* clade. This suggests horizontal transfer from *D. melanogaster* to *D. yakuba*, prior to the transfer from *D. yakuba* to *D. erecta*. If this is the case, these transfers would both have to have occurred very recently, as the smallest divergence between elements in *D. melanogaster* and *D. yakuba* is extremely small, at 0.007. Roo is therefore an interesting case of multiple horizontal transfer events occurring involving the same family and many different species combinations.

Rover has a more restricted, and frequent distribution than Roo, present in the close relatives *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and

D. erecta. Divergence is smaller than for *Adh* for all comparisons involving *D. erecta*, with the exception of *D. simulans*, for which the smallest divergence observed is 0.047, compared with 0.045 for *Adh*. The smallest divergence of elements from its close relative *D. sechellia* from *D. erecta* is 0.039. This difference in values may simply be attributed to copy number, which will be discussed later, which poses a further problem in using the cut-off of the divergence between copies of a particular host gene for inferring horizontal transfer. In this case, the most parsimonious explanation for the divergence data is that a single horizontal transfer event involving *D. erecta* occurred, and that the high identity of elements in the other species with *D. erecta* is an artefact of their close relationships with each other. If this is the case, the fact that the smallest divergence between *D. sechellia* and *D. erecta* is smaller than *Adh*, and between *D. simulans* and *D. erecta* is not, is completely uninformative. This illustrates a limitation in using this method to infer horizontal transfer, and the clear need for additional information, such as phylogenetic incongruence, to both support the argument and elucidate a potential direction of transfer, to build a confident case for horizontal transfer. Additional evidence can also help to build a case for a limited number of transfer events, resulting in small divergence observed between related species as a result of their relationships. Transfer from *D. yakuba* to *D. erecta* is strongly supported as this interspecies comparison yields the smallest divergence, of 0.006. Transfer from *D. erecta* to *D. melanogaster* is supported by the phylogenies, in which elements from *D. melanogaster* are found within a clade of *D. erecta* elements. Horizontal transfer also appears to have introduced Rover into the ancestor of *D. simulans* and *D. sechellia* from the *D. erecta* lineage, as all elements from these species group together within the *D. erecta* clade. Previous work has, however, suggested the direction of this transfer may have involved *D. erecta* as the recipient (Vidal et al. 2009). As expected from the incredibly small divergence between elements, incongruence is also observed involving *D. yakuba* and *D. erecta*. *D. yakuba* elements fall within a cluster of *D. erecta* elements (Figure 4.18), suggesting transfer involving *D. erecta*, once again, as the donor species. As with Roo, it appears that in this case, multiple transfers have occurred.

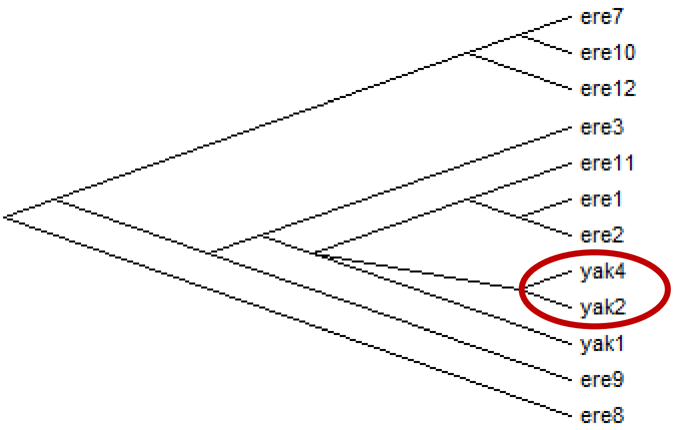


Figure 4.18: Section of the maximum parsimony phylogeny of Rover elements, showing elements from *D. yakuba* falling within a clade of elements from *D. erecta*.

Tabor_DA has an unusual distribution, found only in *D. ananassae* and *D. willistoni*. For such a distribution to be explained under the null hypothesis of vertical transmission, the Tabor_DA family must have been independently lost at least twice: from the ancestor of *D. erecta*, *D. melanogaster* and their relatives, and from the ancestor of *D. persimilis* and *D. pseudoobscura*. Divergence between elements in the two species is not strikingly small, but is much smaller than would be expected given the divergence time separating the two species. Smallest divergence observed is 0.126, compared with 0.225 between the *D. ananassae* and *D. willistoni Adh* genes. Therefore, both small divergence and a patchy distribution support the hypothesis of horizontal transfer having introduced this family into one or other of the two species in which it is present. The phylogenies produced, however, go no way to further support this hypothesis. The trees are congruent with the host relationships in that all elements from *D. ananassae* cluster together to the exclusion of all elements from *D. willistoni*; elements from each species are not dispersed throughout the tree. Clearly whether or not one clade falls within the other would depend upon the rooting of the tree, which cannot be determined. If horizontal transfer introduced the family into a species in which it was previously absent, with only two species on the tree, the phylogenies observed are consistent with expectations. Transfer into a recipient which already contained a resident population of Tabor_DA elements would most likely have resulted in elements from both species being scattered throughout the tree, which is not observed. There is therefore good evidence to support horizontal transfer of Tabor_DA between *D. ananassae* and *D. willistoni* introducing the family into a naïve genome, however, the direction of transfer cannot be inferred.

The **Tirant** family follows the same distribution in the Sophophora as Rover, but unlike Rover, does not represent a particularly striking case of horizontal transfer. Interestingly, there is an insertion of Tirant near to the *parp* locus in a population of *D. simulans*, which might implicate Tirant in the regulation of other transposable element families (Fablet et al. 2009). Divergence between elements in different species is only smaller than *Adh* in two cases, comparison of *D. melanogaster* with its two closest relatives, *D. simulans* and *D. sechellia*. Even in these two cases, the divergence is not strikingly small, with smallest values of 0.018 and 0.012, respectively. Furthermore, the hypothesis of horizontal transfer involving these species is not supported by the phylogenies produced, whereby all elements from *D. melanogaster* form a monophyletic clade, with the *D. simulans*/*D. sechellia* clade as a sister group. However, phylogenetic incongruence is observed, with elements from *D.*

yakuba and *D. erecta* grouping together on the tree, within the clade of elements from *D. simulans* and *D. sechellia*. Tirant is therefore an ambiguous case, with no strong evidence to support horizontal transfer involving any particular pair of species. Transfer from the ancestor of *D. simulans* and *D. sechellia* into the ancestor of *D. yakuba* or *D. erecta* would explain the relationships observed on the tree, however, this is not supported by the divergence data, and would not be possible given the divergence time between *D. yakuba* and *D. erecta*. Were this transfer relatively ancient, this may lead to divergence exceeding that of the constrained host gene *Adh*.

Transpac has a relatively wide distribution in the Sophophora, found in the same species as Tirant and Rover, in addition to *D. ananassae*. However, horizontal transfer is only inferred to have occurred between the closest relatives among these six species, *D. melanogaster*, *D. simulans* and *D. sechellia*. All three interspecies comparisons involving these species result in strikingly small divergence, of 0.000 for *D. melanogaster*/*D. simulans*, and 0.001 for the other two comparisons. These data suggest multiple horizontal transfer events have occurred recently. Transfer between *D. melanogaster* and the ancestor of *D. simulans* and *D. sechellia* would not be expected to yield such small divergence, and therefore at least two separate transfers involving *D. melanogaster* can be inferred. A third transfer, between *D. simulans* and *D. sechellia*, is also likely. This final transfer cannot be supported by evidence from the phylogenies produced, however, the remaining two transfers are supported by phylogenetic incongruence. The single element from *D. simulans* is found within the clade of *D. melanogaster* elements, suggesting *D. simulans* was the recipient species of this extremely recent transfer. This suggests that Transpac would have been stochastically lost from the *D. simulans* lineage had horizontal transfer not occurred, and indicates the importance of horizontal transfer to the survival of transposable element families in *Drosophila*. Although elements from *D. sechellia* do not fall within the *D. melanogaster* clade, many of them are more closely related to the elements from *D. melanogaster* than the other elements from their own species, again suggesting transfer with *D. melanogaster* as the donor. The divergence among all elements in *D. sechellia* and *D. melanogaster* is extremely small, therefore it is impossible to be confident about the role of *D. melanogaster* as donor in this transfer. In both species, all elements present are closely related to each other and appear to be recent insertions. It is possible that once again this represents a case of a family having been lost from a particular lineage and reintroduced by horizontal transfer.

There is one family of LTR retrotransposons distributed throughout the Sophophora but absent from the Drosophila subgenus for which horizontal transfer is only suggested by the observation that the smallest divergence between elements of the same family in different species is less than that of the host gene *Adh*.

Accord2 is found in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta* and *D. ananassae*. Although they share a name, alignment between Accord and Accord2 is relatively poor, for example compared with the Gypsy families, with around 55% identity between the two consensus sequences. The distribution of Accord2 is consistent with vertical transmission and the position of *D. ananassae* elements in the phylogenies produced is congruent with known host relationships. However, small divergence of elements from *D. melanogaster*, *D. sechellia*, *D. yakuba* and *D. erecta*, when compared with elements from *D. ananassae*, suggests a horizontal transfer event involving *D. ananassae* may have occurred.

There are six families distributed throughout the Sophophora, but absent from the Drosophila subgenus, for which divergence between elements in different species does not suggest that horizontal transfer has occurred. These families are **Gypsy7**, **Gypsy8**, **Idefix**, **Invader5**, **Nobel** and **Tram**. It is likely that the distribution of these families throughout the Sophophora can be attributed to vertical transmission. The distribution of Gypsy7, Idefix and Tram is consistent with vertical transmission. The distribution of the remaining families is patchy, but could be attributed to stochastic loss. Gypsy8 appears to have been deleted from *D. sechellia*, and Invader5 from *D. yakuba*. Loss of Gypsy8 from *D. sechellia* is confirmed by the presence of solo LTRs belonging to this family in this species. Nobel has an unusual distribution, found in *D. yakuba*, *D. ananassae*, *D. pseudoobscura* and *D. persimilis*. For stochastic loss to explain this distribution, Nobel must have been deleted from *D. erecta* and from the ancestor of *D. melanogaster*, *D. simulans* and *D. sechellia*. Alternatively, horizontal transfer of Nobel may have occurred long into the past, with mutations accumulating since to the extent that the event is no longer detectable.

4.3.4.4 TV1, an LTR retrotransposon shared by *D. virilis* and *D. mojavensis* only

Due to the lack of complete sequenced genomes from the Drosophila subgenus, with three sequenced representatives compared with nine of the Sophophora, and the relative lack of study of the transposable elements of these species, only one family, TV1, was implicated to have been involved in horizontal transfer and found only in species of the Drosophila subgenus. **TV1** is found in *D. mojavensis*, and its closest relative amongst the species for which the complete genome sequence is available, *D. virilis*. Therefore, the distribution is consistent with vertical transmission. The smallest divergence between the single element in *D. mojavensis*, and the nine elements in *D. virilis*, is 0.009, compared with a divergence of 0.144 in the *Adh* gene between these species, which diverged around 24 million years ago. Divergence ranges between 0.009 and 0.012, suggesting recent horizontal transfer. However, as only one element is present in *D. mojavensis*, phylogenetic incongruence cannot be observed, and a potential direction of transfer cannot be inferred. It is perhaps likely, given that all elements in *D. virilis* are closely related to the element from *D. mojavensis*, that a recent horizontal transfer event introduced this family to one or other of these two species. It is also possible that the family was present in both species in the past, but more ancient elements have been deleted from the genomes.

4.3.4.5 LTR retrotransposon families distributed throughout the Drosophila and Sophophora subgenera

As transposable elements are rapidly eliminated from *Drosophila* genomes, families surviving to be present in both species of the Sophophora and the Drosophila subgenera, which diverged around 40 million years ago, are relatively rare. For the reasons discussed in the section above, knowledge of transposable elements, and the amount of sequence data available, is limited for the Drosophila subgenus. It is therefore, perhaps, given only one family was limited to the Drosophila subgenus (TV1, see above), surprising that, although a small fraction of the total number of LTR retrotransposon families in the genus, so many families are found distributed across both the Drosophila and Sophophora subgenera. The possibility that those which are found in both subgenera have survived due to horizontal transfer is discussed below.

Horizontal transfer of eleven families distributed across the *Drosophila* genus is supported by small divergence, phylogenetic incongruence and a patchy distribution across the host phylogeny.

Gtwin is distributed across both the Sophophora and Drosophila subgenera, but is present in six species of the Sophophora (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta* and *D. ananassae*) and only one of the Drosophila, *D. mojavensis*. Within the Sophophora, the distribution of Gtwin is consistent with vertical transmission and suggests that the family was present in the ancestor of these six species and has since been inherited from this ancestor and retained in all lineages. However, the presence of Gtwin in a single species of the Drosophila subgenus may suggest that horizontal transfer has occurred, introducing Gtwin into either *D. mojavensis* at a time since its split from its closest relative for which the sequenced genome is available, *D. virilis*, or into the ancestor of the six Sophophoran species in which the family is present. However, the divergence between elements does not support such a hypothesis, as the smallest divergence values, although small, are not smaller than for the host gene *Adh* between these species. Only a single Gtwin element is present in the *D. mojavensis* genome sequence, and therefore its position on the phylogenies produced would depend upon rooting of the tree, and therefore cannot be used to make inferences about its relationships with the other elements on the trees. Therefore, in the case of Gtwin, horizontal transfer between the two subgenera is only supported by the patchy distribution of the family across the *Drosophila* phylogeny, which could also be explained by the presence of

Gtwin in the ancestor of the subgenera around 40 million years ago, followed by stochastic loss on four lineages, leading to the following species: *D. pseudoobscura*/*D. persimilis*, *D. willistoni*, *D. virilis* and *D. grimshawi*, for which complete genome sequences are available. Although transfer of Gtwin across subgenera is not well supported, transfer within the Sophophora is supported by both small divergence between elements in different species and by phylogenetic incongruence (Figure 4.19). Divergence is smaller than for the host gene *Adh* for four interspecies comparisons, that of *D. melanogaster* with *D. simulans*, *D. sechellia* and *D. erecta*, and also between *D. yakuba* and *D. erecta*. Smallest divergence values for these comparisons were 0.013, 0.012, 0.008 and 0.035, respectively. Gtwin has been reported to have horizontally transferred into *D. erecta* (Kotnova et al. 2007).

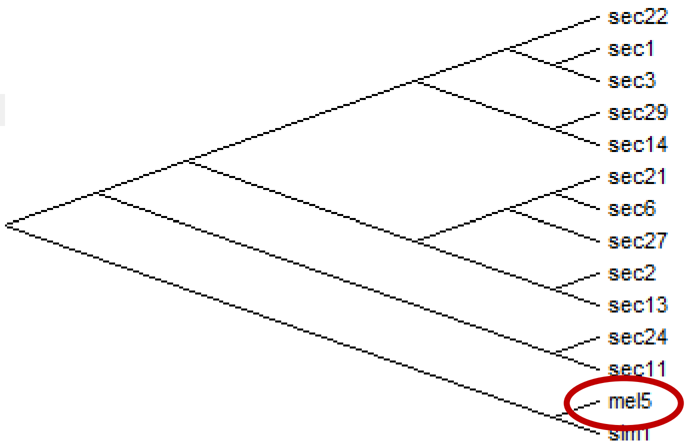


Figure 4.19: Section of the maximum parsimony phylogeny of Gtwin elements, showing a single element from *D. melanogaster* falling within a clade of elements from *D. simulans* and *D. sechellia*.

Gypsy6 is an interesting example of an LTR retrotransposon family which follows a patchy distribution. The family is distributed across both the Drosophila and Sophophora subgenera, found in *D. virilis* and *D. mojavensis* of the former subgenus, and *D. melanogaster*, *D. yakuba* and *D. erecta* of the latter. Its distribution is unusual in two respects. Firstly, it is present in both subgenera, perhaps suggesting its presence in the ancestor of the two subgenera 40 million years ago, however, it is absent from numerous descendent lineages. Secondly, the family is absent from *D. simulans* and *D. sechellia*, despite its presence in the closely-related species *D. melanogaster*. The family must have been lost at least five times to account for its distribution in the absence of horizontal transfer (*D. simulans*/*D. sechellia*, *D. ananassae*, *D. pseudoobscura*/*D. persimilis*, *D. willistoni* and *D. grimshawi* lineages). However, horizontal transfer introducing this family into either the Drosophila or the Sophophora from the other subgenus is supported neither by the divergence data nor the phylogenetic trees. Divergence between elements in intersubgenus interspecies comparisons is always greater than for the host gene *Adh*, and *D. virilis* and *D. mojavensis* elements cluster together on the trees, to the exclusion of all other elements. However, this is to be expected were the horizontal transfer event to have occurred quite some time in the past (prior to the divergence of *D. virilis* and *D. mojavensis*), and therefore does not rule out the possibility of horizontal transfer between the subgenera. Divergence is smaller than for *Adh* for four interspecies comparisons, three within the Sophophora (*D. melanogaster*/*D. yakuba*, *D. melanogaster*/*D. erecta*, and *D. yakuba*/*D. erecta*), and one within the Drosophila (*D. virilis*/*D. mojavensis*). Transfer between *D. virilis* and *D. mojavensis* is supported, with one element from *D. virilis* more closely related to the element from *D. mojavensis* than the other elements from *D. virilis*, suggesting transfer involving *D. mojavensis* as the donor species. This relationship is well-supported and consistent across both phylogenetic construction methods. Incongruence involving *D. melanogaster* is not observed.

In the case of **HMS Beagle**, the distribution of the family is patchy due to the absence of the family from certain lineages, perhaps suggesting horizontal transfer introduced HMS Beagle into certain groups. The family is found in the close relatives *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta* of the Sophophora, but is absent from the other Sophophoran species. It is also found in *D. virilis*, and its closest relative for which the sequenced genome is available, *D. mojavensis*, in the Drosophila subgenus. This distribution is more consistent with horizontal transfer than stochastic loss, as it would require the family to be present in the ancestor of the Sophophora and Drosophila, and to have been independently lost from at least four lineages (*D. ananassae*, *D. pseudoobscura*/*D. persimilis*, *D. willistoni* and *D. grimshawi*). The presence of the family in two groups of closely-related species found in distinct parts of the tree gives the impression of horizontal transfer introducing the family into one or other of the groups. However, such a hypothesis is not supported by divergence data, as when comparing HMS Beagle elements from Sophophoran species and Drosophila species, divergence is consistently higher than for the host gene *Adh*. This could, however, be explained by a relatively ancient transfer event, which would be required to introduce the family into either the ancestor of the two Drosophila species, or the five Sophophoran species, in which the family is present. Therefore this may not be particularly informative in inferring whether or not transfer over this distance has occurred. No further evidence for such a transfer is found from the phylogenies produced, in which elements from *D. virilis* and *D. mojavensis* group together to the exclusion of the elements from other species. Again, this is to be expected given the inferred transfer would have occurred before the radiation of the species present on the tree. Horizontal transfer between the two subgenera, in this case, is possible but not supported by a large amount of evidence. However, horizontal transfer of

HMS Beagle within the Sophophora is supported by two lines of evidence. Divergence is smaller than for *Adh* in comparison of elements from *D. melanogaster* and *D. simulans* (0.011), and between *D. yakuba* and *D. melanogaster* (0.000), *D. simulans* (0.004) and *D. sechellia* (0.019). This is incredibly striking, with a pair of elements in *D. melanogaster* and *D. yakuba* sharing 100% identity, which, given the divergence time between these two species, is strong evidence for more recent common ancestry, and therefore horizontal transfer. Such a transfer is supported by the phylogenies produced, as three elements from *D. yakuba* fall within, or just outside, the *D. melanogaster* clade, such that elements from *D. yakuba* are more closely related to *D. melanogaster* elements than are those from *D. simulans* and *D. sechellia*. This supports a direction of transfer involving *D. melanogaster* as the donor and *D. yakuba* as the recipient. Furthermore, elements from *D. simulans* and *D. sechellia* fall outside the *D. melanogaster*/*D. yakuba* clade, rather than forming a discrete group alone. This might suggest horizontal transfer between *D. melanogaster* and *D. simulans*, *D. sechellia* or the ancestor of these two species. If this is the case, such a transfer along with the transfer from *D. melanogaster* to *D. yakuba* would be sufficient to explain the small divergence between elements in the four interspecies comparisons. Two possible routes are possible. The first is that a transfer occurred from *D. simulans* into *D. melanogaster*, and then that an element descended from this transfer event in *D. melanogaster* transferred into *D. yakuba*. The second is that related elements transferred from *D. melanogaster* into both *D. yakuba* and *D. simulans*. The first of these hypotheses is more likely, as this would explain the high identity between elements in *D. yakuba* and *D. sechellia*, which can be attributed to the close relationship between *D. simulans* and *D. sechellia*, such that elements in *D. sechellia* would resemble those that transferred from *D. simulans*.

Invader3 is an interesting case, as although three separate lines of evidence support horizontal transfer of this family, they do not provide evidence to support the same transfer. Invader3 has a patchy distribution, as it is absent from several lineages. The family is found in all species of the Sophophora except for *D. ananassae* and *D. willistoni*, and is only found in *D. virilis* of the Drosophila subgenus. This unusual distribution may be explained by horizontal transfer into *D. virilis*, however this is supported neither by the divergence between elements, nor by the phylogenies, in which elements from *D. virilis* are monophyletic and would form a congruent root for the trees. Absence from two species of the Sophophora is most likely explained by stochastic loss. Divergence between Invader3 elements in different species is only smaller than for *Adh* in the case of *D. melanogaster* and *D. simulans*, however, once again, incongruence involving these two species is not observed on the tree. The only example of phylogenetic incongruence observed on the tree is between *D. melanogaster* and *D. sechellia*, which is consistent and well-supported. A single element from *D. sechellia* falls into the *D. melanogaster* clade, suggesting transfer involving *D. sechellia* as the recipient (Figure 4.20). The entire clade of *D. melanogaster* elements is found with a group of *D. sechellia* elements as its closest relatives, rather than these elements being found with the others from *D. simulans* and *D. sechellia*. This supports the hypothesis of a horizontal transfer event involving *D. melanogaster* as the donor. This direction of transfer does not explain the small divergence between *D. melanogaster* and *D. simulans*. The divergence between *D. melanogaster* and *D. sechellia* is only slightly in excess of that for *Adh*, therefore a horizontal transfer event involving these two species is best supported for Invader3.

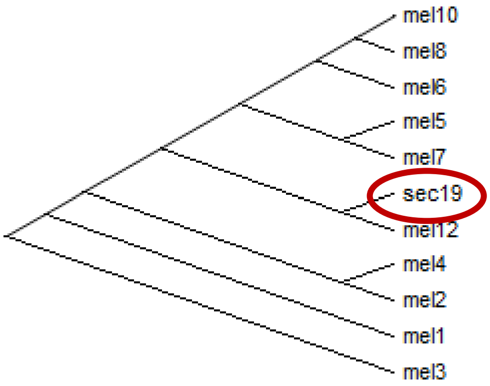


Figure 4.20: Section of the maximum parsimony phylogeny of Invader3 elements, showing a single element from *D. sechellia* falling within a clade of elements from *D. melanogaster*.

The distribution of **Invader6** is patchy, and may suggest horizontal transfer occurring between the Sophophora and Drosophila. The family is present in the close relatives *D. melanogaster*, *D. simulans*, *D. sechellia* and *D. yakuba*, and their distant relative *D. mojavensis*. However, the divergence between elements from the Sophophoran species and *D. mojavensis*, although small given the divergence time between the species, is not smaller than the divergence between the *Adh* genes. However, due to the extent of the divergence between the species, it is possible that a horizontal transfer event involving the ancestor of *D. mojavensis* occurred a considerable amount of time in the past, before the divergence of *D. melanogaster* and *D. yakuba*, which would account for the relatively small amount of divergence between elements. The phylogenies produced do not provide evidence to support a transfer between the two subgenera, but this is to be expected were such a transfer to have occurred before the divergence of the Sophophoran species. Divergence between elements in different species is smaller than *Adh* for all comparisons involving *D. melanogaster* and all comparisons involving *D. yakuba*, however, horizontal transfer involving *D. yakuba* is not supported by the phylogenies, whereby all elements from *D. yakuba* cluster together in the appropriate location on the trees. However, transfer from *D. simulans*, *D. sechellia*, or the ancestor of these two species into *D. melanogaster* is supported by the trees. Such a transfer would explain the small divergence between elements in *D. melanogaster* and both *D. simulans* and *D. sechellia*. The divergence between elements suggests incredibly recent transfer involving *D. yakuba* (ranging from 0.005 compared with *D. sechellia*, to 0.006 compared with *D. melanogaster* and *D. simulans*). However, a recent transfer, which would be likely to have introduced Invader6 into *D. yakuba* to account for the clustering of these elements together on the tree, and the absence of Invader6 from *D. erecta*, would not explain the congruence of the position of *D. yakuba* in the Invader6 tree compared with the host tree.

Micropia is distributed across the *Drosophila* genus in the same species as HMS Beagle: *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta* from the Sophophora, and *D. mojavensis* and *D. virilis* from the Drosophila. Therefore, once again, this patchy distribution is consistent with a possible horizontal transfer between the ancestor of *D. mojavensis* and *D. virilis*, and the ancestor of the five Sophophoran species. However, once again, such a hypothesis is not supported by the divergence data. The phylogenies are congruent with respect to the host phylogeny in terms of the relationship between the subgenera, with all elements from *D. virilis* and *D. mojavensis* grouping together to the exclusion of all elements from the other five species, however, this would be expected both under the hypothesis of horizontal transfer proposed, and also under vertical transmission. Divergence is only smaller than the host genes for all three interspecies comparisons involving the close relatives *D. melanogaster*, *D. simulans* and *D. sechellia*. Smallest divergence observed is much smaller than for *Adh*, ranging from

0.004 between *D. simulans* and *D. sechellia*, to 0.005 for both comparisons with *D. melanogaster*. Transfers involving these species are supported by incongruent phylogenies, however, individual relationships are inconsistent between the two trees, most likely due to the lack of mutation data available to deduce relationships. However, all elements from *D. melanogaster* are consistently found within the *D. simulans*/*D. sechellia* clade. This would suggest horizontal transfer occurring with the *D. simulans*/*D. sechellia* ancestor as the donor, and *D. melanogaster* as the recipient, which is consistent with the divergence observed. Transfer between *D. simulans* and *D. sechellia*, as in all cases, cannot be reliably inferred from the tree, and is not strongly supported by the divergence data, which due to the extremely close relationship between these two species, may be smaller than *Adh* simply by chance. Further incongruence is observed on the tree, for example between *D. simulans* and *D. yakuba*. Such a transfer is not supported by divergence data and may be explained by differential retention of an ancestral polymorphism. For this reason, phylogenetic incongruence alone does not provide sufficient evidence to infer horizontal transfer.

Ninja has an extensive distribution, present in nine of the twelve sequenced *Drosophila* genomes. The family is present in all Sophophoran species with the exception of *D. willistoni*, in addition to *D. virilis* from the Drosophila subgenus. This patchy distribution would be consistent with vertical transmission within the Sophophora and horizontal transfer from a Sophophoran species into *D. virilis*. Nine interspecies comparisons yield divergence smaller than that of *Adh*, including comparisons of *D. melanogaster* and *D. sechellia* elements with those from *D. virilis*. However, if the phylogenies produced are rooted on *D. virilis* (under the null hypothesis of vertical transmission), the trees are congruent with respect to the position of the main groups, although there is some inconsistent incongruence within the *melanogaster* group. All elements from *D. virilis* form a monophyletic clade on the tree, however, this is as would be expected if horizontal transfer had introduced the family to this species. Horizontal transfer of Ninja is supported within the *melanogaster* group, with small divergence in comparison of elements from *D. melanogaster*, *D. sechellia* and *D. yakuba*, and consistent incongruent relationships. For example, the element designated mel5 clusters with sec9 and sec5, and yak3 and yak7 cluster with mel4 and mel7. The trees of the LTRs also support horizontal transfer involving these species.

Osvaldo has the most extensive distribution across the twelve sequenced *Drosophila* genomes of the LTR retrotransposon families, present in ten of the twelve species. This extensive distribution is only seen for one other family, Stalker4. Osvaldo is absent from one Sophophoran species, *D. willistoni*, and one Drosophila species, *D. virilis*. Divergence between elements in different species is not strikingly small, and in fact is consistent with vertical transmission in all but one case. This suggests that the absence of the family from *D. willistoni* and *D. virilis* is due to stochastic loss. One interspecies comparison, that of *D. ananassae* and *D. grimshawi*, yields a smallest divergence less than that of *Adh*, with a value of 0.171, compared with 0.252 for *Adh*. If this is due to horizontal transfer, this would represent a transfer between the two subgenera. Osvaldo is one of only two LTR retrotransposon families found in *D. grimshawi* along with other species (Gypsy_DG is found in *D. grimshawi* but is restricted to this species). Interestingly, Uhu, described previously, the only DNA transposon family found in *D. grimshawi* in addition to other species, is also implicated in transfer with *D. ananassae*. The two species share geographical overlap in Hawaii (Ashburner and Novitski 1976). Neither of the two proposed transfer events involving these species appear to be recent, and in this case, although the smallest divergence is less than that of *Adh*, it is not particularly small. In the case of Osvaldo, the patchy distribution observed does not provide further evidence for transfer involving *D. ananassae* and *D. grimshawi*. Furthermore, the phylogenetic incongruence observed does not involve these species. *Drosophila grimshawi* elements form a monophyletic clade, with the clade of all elements from *D. mojavensis*, the other species of the Drosophila subgenus in which Osvaldo is found, the sister clade. In addition, all elements from *D. ananassae* form a monophyletic clade, however, the position of this clade in the tree is consistently incongruent, forming a clade with elements from *D. persimilis* and *D. pseudoobscura*, rather than the elements found in species of the *melanogaster* group. Further examples of incongruence, involving the species of the *melanogaster* group, are observed, however these are inconsistent between the trees produced using the two phylogenetic construction methods.

RooA has an unusual distribution, as it is found in the five closely-related members of the *melanogaster* group (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*), in addition to the distant relative from the Drosophila subgenus, *D. mojavensis*. Most RooA elements appear to be inactive, but are able to propagate through the action of helper elements (de la Chaux and Wagner 2009). Divergence is smaller than for *Adh* for three interspecies comparisons within the Sophophora, all involving *D. erecta*. Divergence between elements therefore does not support the hypothesis of horizontal transfer as an explanation for the patchy distribution of this family, found in five close relatives among the Sophophora and absent from the relatives of *D. mojavensis*. Furthermore, the phylogenies produced do not support such a transfer, as all elements from *D. mojavensis* cluster together on the tree, to the exclusion of all other elements. It is possible that the RooA family has been vertically transmitted and has survived in both subgenera for over 40 million years, having been lost from numerous lineages during that time. Alternatively, transfer may have occurred involving an unknown donor species, either into *D. mojavensis* or into the ancestor of the *melanogaster* group. Beyond that, there is the possibility that horizontal transfer between the two subgenera occurred, but too far into the past to be detected. If transfer had involved the ancestor of the *melanogaster* group and the ancestor of *D. mojavensis*, this would not be so ancient as to expect the presence of RooA in *D. virilis*, but would be sufficient time that mutations would accumulate such that the divergence exceeded that of the constrained host gene *Adh*. Additionally, transfer of this manner would result in a phylogeny congruent with known host relationships. Within the Sophophora, divergence between elements in *D. simulans*, *D. sechellia* and *D. yakuba*, when compared with *D. erecta*, is smaller than for *Adh*. Phylogenetic incongruence involving these species is also observed. Elements from *D. erecta* and *D. yakuba* consistently cluster together, but fall in two places on the tree. One location is congruent with the host phylogeny, whereas the other cluster falls within the clade of elements from *D. simulans*, *D. sechellia* and *D. melanogaster*. This interesting series of relationships might imply transfer from the ancestor of *D. simulans*, *D. sechellia* and *D. melanogaster* into either *D. yakuba* or *D. erecta*, or their ancestor. RooA is therefore an interesting case whereby a large amount of incongruence in the phylogeny and multiple examples of small divergence can be explained by a limited number of horizontal transfer events.

Stalker4, along with Osvaldo, has the most extensive distribution observed for any LTR retrotransposon family in the *Drosophila* genus. It is found in ten of the twelve species for which the sequenced genomes are available, absent only from *D. ananassae* of the Sophophora, and *D. grimshawi* of the

Drosophila subgenus. Its distribution therefore differs from that of Osvaldo. The absence of Stalker4 from *D. ananassae* is most likely explained by stochastic loss from the lineage leading to this species, as the family is present in all of the closest relatives of *D. ananassae*, and its more distant relatives in the Sophophora. However, the presence of Stalker4 in *D. mojavensis* and *D. virilis*, and its absence from *D. grimshawi* might be explained either by stochastic loss from the lineage leading to *D. grimshawi*, or by horizontal transfer into the ancestor of *D. virilis* and *D. mojavensis*, or each of these species individually, from the Sophophora. Stalker4 is an interesting family to investigate partly due to its extensive distribution, but also due to the interesting relationships between elements in different species, which support horizontal transfer between the two subgenera. Stalker4 is one of only four LTR retrotransposon families for which transfer between the Drosophila and Sophophora subgenera is supported by divergence data. Interestingly, as such an event is apparently so rare, there is evidence to suggest that transfer of Stalker4 between species of the two different subgenera has occurred more than once. There are 24 interspecies comparisons, the greatest number for any LTR retrotransposon family, for which the divergence between elements can be smaller than that of the host gene *Adh*. These divergence data are likely to be explained by relatively few horizontal transfer events. Divergence can be less than that of *Adh* both for interspecies comparisons within the Sophophora, and also between the Sophophora and Drosophila subgenera, but not within the Drosophila subgenus, where the extent of divergence between *D. virilis* and *D. mojavensis* is consistent with vertical transmission. The divergence data do not support a single transfer between the Drosophila and Sophophora subgenera involving the ancestor of *D. melanogaster* and *D. willistoni* and the ancestor of *D. mojavensis* and *D. virilis*. The divergence between both *D.*

pseudoobscura and *D. persimilis* compared with *D. virilis* is much smaller than would be expected under this scenario, at only 0.059 and 0.052, respectively. These data support the hypothesis of horizontal transfer of Stalker4 from the ancestor of *D. pseudoobscura* and *D. persimilis* into *D. virilis*. The divergences of elements from the other Sophophoran species from *D. virilis* are, however, smaller than would be expected under this scenario, supporting further transfers between *D. virilis* and the Sophophora. Additionally, it appears that horizontal transfer may have occurred between *D. willistoni* and *D. mojavensis*, most likely with the latter as the donor, as the identity of elements from the other Sophophoran species with elements from *D. mojavensis* is relatively small.

Putative transfers within the Sophophora may go some way to explain the unusual divergence between elements of the Sophophora and the Drosophila species in the absence of more than two horizontal transfer events. There is evidence to suggest transfer occurring between *D. pseudoobscura*, *D. persimilis* or their ancestor and other members of the Sophophora, as well as transfer involving *D. willistoni* and other members of the subgenus. If these transfers involved elements which were related to those elements involved in the transfers between the subgenera, this would explain the unexpectedly high identity between, for example, the *melanogaster* group and *D. virilis*, had the transfer into *D. virilis* originated in the ancestor of *D. persimilis* and *D. pseudoobscura*. However, the precise transfers that may have occurred within the Sophophora are difficult to elucidate. Divergence between elements in different Sophophoran species is not strikingly small in any particular case, and although phylogenetic incongruence is observed for these species, relationships are inconsistent between the two phylogenetic reconstruction methods, and do not involve the key species *D. pseudoobscura*, *D. persimilis* and *D. willistoni*. The phylogenies do support both putative transfers between the subgenera, with the elements of *D. mojavensis* falling within the *D. willistoni* clade, supporting *D. willistoni* as the donor species. However, contrary to expectations, the elements from *D. pseudoobscura* and *D. persimilis* fall within the *D. virilis* clade on the maximum parsimony tree, not the other way around. This would imply *D. virilis* was the donor species, and therefore would not explain the high identity between elements from other Sophophoran species and *D. virilis*. However, the maximum likelihood tree presents the two groups as sister taxa, consistent with introduction of Stalker4 into *D. virilis* from the *D. pseudoobscura*/*D. persimilis* ancestor. As the two trees are inconsistent with respect to this relationship, the direction of transfer cannot be conclusively determined, however, *D. virilis* as the recipient species is more consistent with the divergence data observed.

Zam is distributed across the *Drosophila* genus, found in six of the twelve species with sequenced genomes available. Interestingly, the Zam family is believed to be regulated by the activity of the COM locus in *Drosophila* (Bergman et al. 2006), which may limit its proliferation, or apply selection pressure for Zam to invade other species. Zam follows the distribution of RooA, found in the *melanogaster* group and the distant relative of those species, *D. mojavensis*. Similarly to RooA, it is possible that Zam was introduced into *D. mojavensis* from a species of the Sophophora subgenus. However, the divergence between elements is only smaller than the host gene *Adh* for a single interspecies comparison, that of *D. yakuba* and *D. erecta*, within the Sophophora. Therefore divergence between elements does not support the idea that the patchy distribution of this family is due to horizontal transfer. Furthermore, transfer into the Drosophila subgenus is not supported by the phylogenies produced, in which all elements from *D.*

mojavensis form a clade to the exclusion of all other elements. Therefore, the same possible explanations for the unusual distribution of this family apply as in the case of RooA. Horizontal transfer of Zam between *D. yakuba* and *D. erecta* is supported through small divergence. The clustering of elements from the two species on the trees is consistent between both phylogenetic reconstruction methods, but this relationship is congruent with the host phylogeny. The smallest divergence observed between elements in these species is only 0.011, considerably smaller than the divergence of 0.051 between their *Adh* genes.

Horizontal transfer of a further LTR retrotransposon family distributed across the *Drosophila* phylogeny, **GypsyDS**, is supported by both small divergence and a patchy distribution. The family is present in *D. persimilis* of the Sophophora and *D. virilis* of the Drosophila subgenus. The two phylogenies produced are consistent and congruent with the host phylogeny, with all elements from *D. persimilis* clustering together to the exclusion of all elements from *D. virilis*. The presence of the family in only a single species of each subgenus strongly suggests horizontal transfer has occurred. The lack of the family from *D. pseudoobscura*, which diverged from *D. persimilis* only around two million years ago, is particularly interesting, and may suggest transfer occurred from *D. virilis* to *D. persimilis*. However, deletion of elements from *D. pseudoobscura* appears to be particularly frequent. Therefore, in this case, a direction of transfer cannot be confidently inferred. However, smallest divergence of only 0.052 (range 0.052-0.104), compared with 0.221 for *Adh* between these species, strongly supports the hypothesis of horizontal transfer.

There are no LTR retrotransposon families distributed across the Sophophora and Drosophila subgenera for which horizontal transfer is supported by the single piece of evidence that the smallest amount of divergence between two elements of the same family in different species is less than for *Adh*. The case for horizontal transfer is supported in the majority of cases by a further two pieces of evidence.

There are two families, **Copia2** and **Invader2**, which are distributed throughout the *Drosophila* genus, but divergence between elements in different species does not suggest horizontal transfer has occurred. It is possible that these families have managed to survive to be transmitted vertically, or, alternatively, horizontal transfer may have occurred in the relatively distant past, such that it is no longer detectable due to a long period of mutation. In spite of this potential problem in the detection of horizontal transfer for such distantly-related groups as the Sophophora and Drosophila subgenera, only two families provide no evidence for horizontal transfer, compared with 6 families restricted to *D. melanogaster*, *D. simulans* and *D. sechellia*, for which any transfer would be recent. This could of course be due to a longer time for any transfer to occur, and a greater number of pairs of species between which transfer could occur. Both Copia2 and Invader2 have a patchy distribution across the phylogeny, with the pattern of their absence from certain lineages suggesting stochastic loss.

4.3.5 Geographical distribution

One of the prerequisites which must be accounted for before horizontal transfer can be confidently inferred is geographical overlap between the donor and recipient species. Although there may be some evidence supporting the hypothesis of horizontal transfer, if the two species inferred to have been involved do not come into contact with each other, it can be assumed that there would be no manner by which genetic material could be transferred between them. Therefore, for each putative case of horizontal transfer described above, it was determined whether or not there is overlap in the geographical range of the inferred donor and recipient species.

Species	**Distribution**				
	Africa	North America	South America	Oriental	Europe
D. melanogaster	🟦	🟦	🟦	🟦	🟦
D. simulans	🟦	🟦	🟦	🟦	🟦
D. sechellia	Seychelles				
D. yakuba	🟦				
D. erecta	Central				
D. ananassae	🟦	🟦	🟦	🟦	🟦
D. pseudoobscura		🟦	Colombia		
D. persimilis		West			
D. willistoni		South, W. Indies	Brazil, Bolivia		
D. virilis		🟦	🟦	🟦	
D. mojavensis		🟦			
D. grimshawi		Hawaii			

Table 4.10: Geographical distribution of the twelve *Drosophila* species. Blue shading indicates the presence of a particular species in the corresponding region. Where a species is only present in part of a region, its distribution is given (Ashburner and Novitski 1976).

	mel	sim	sec	yak	ere	ana	pse	per	wil	vir	moj	gri
mel		38	29	20	9	3			1	2		
sim	●		21	12	1	1			1	1		
sec	●	●		10	2	1			1	1		
yak	●	●	●		13	1					1	
ere	●	●	●	●		1						
ana	●	●	●	●	●				1	2		2
pse	🔵	🔵				🔵		1		2	2	
per	🔵	🔵				🔵	●			3	2	
wil	●	●	🔴			●	🔵	🔵			2	
vir	●	●	🔴			●	●	●	🔵		3	
moj	🔵	🔵		🔴		🔵	●	●	●	●		
gri	🔵	🔵				●	🔵	🔵	🔵	🔵	🔵	

Figure 4.21: Summary of horizontal transfer events for each interspecies comparison. The number of inferred events is given in the top right. Blue shading indicates geographical overlap between the two species. Black dots indicate interspecies comparisons for which horizontal transfer is inferred and geographical overlap is observed. Blue dots indicate pairs of species which overlap geographically but for which there is no evidence of horizontal transfer. Red dots indicate inferred transfers where the putative donor and recipient species do not overlap.

Each inferred horizontal transfer event was examined to determine whether the putative donor and recipient species overlap geographically. Where the ancestor of a group of species is believed to have been involved, this is counted as a transfer involving each of the descendent species, and will therefore overestimate the number of horizontal transfer events. It was determined that for the majority of inferred horizontal transfer events (187/190), there was overlap between the geographical range of the putative donor and recipient species. In the three remaining cases, the identity shared between transposable elements in the two non-overlapping species can be explained by transfers involving others species. The first is the Copia family, for which comparisons of elements in *D. sechellia* and *D. willistoni* are smaller than for the host genes, although these two species do not overlap. This similarity could be an artefact of a transfer of Copia from either *D. melanogaster* or *D. simulans* into *D. willistoni*. The second involves the Ninja family, where small divergence is observed between elements in *D. sechellia* and *D. virilis*. However, once again, this may be a consequence of a transfer from *D. melanogaster* to *D. virilis*. The third case cannot be so easily explained. Elements of the Minos family from *D. yakuba* and *D. mojavensis* share high levels of identity, although these species do not overlap. This family is not present in any of the other ten species for which the complete sequenced genome is available. However, previous studies of Minos have identified horizontal transfer events between *D. mojavensis* and members of the *saltans* group (Arca and Savakis 2000;de Almeida and Carareto 2005). It may be that Minos has been horizontally transferred between *D. yakuba* and *D. mojavensis* via an intermediate species which overlaps with both, therefore indicating at least two transfer events involving this family.

As examination of geographical overlap was conducted following the inference of horizontal transfer, this provides additional support for horizontal transfer as the explanation for the observations made, such as small divergence between elements, rather than other factors. Were other factors, such as constraint, frequently involved, it would be expected that cases of small divergence between elements of the same family in different species would also be observed between species pairs which do not overlap in the geographical range. This could then potentially be attributed to constraint rather than recent common ancestry. However, it may be that constraint is more likely to be observed between closely-related species pairs, which may be more likely to overlap in their geographical range as a result of their recent common ancestry.

4.4 Discussion

Through examination of three lines of evidence: small divergence between transposable elements of the same family in different species, phylogenetic incongruence and patchy distribution across the host phylogeny, it was determined that horizontal transfer has occurred at least once for a large proportion of the transposable element families investigated. This investigation adds further weight to the growing assumption that horizontal transfer is an important part of the lifecycle of transposable elements in *Drosophila* (Bartolome et al. 2009;Loreto et al. 2008), allowing them to continue to survive within genomes which rapidly eliminate them. This is contrary to previous expectations, that horizontal transfer would be found to be a rare event, affecting only a few transposable element families.

Family	Species												Total
	mel	sim	sec	yak	ere	ana	pse	per	wil	vir	moj	gri	
Bari	✓	✓	✓		✓	✓							5
Helitron1	✓					✓							2
Mariner		✓	✓	✓									3
Transib2	✓	✓	✓	✓	✓								5
Hobo	✓	✓	✓	✓	✓								5
TransibN2							✓	✓					2
hAT1N						✓	✓	✓			✓		4
Looper	✓	✓	✓	✓		✓				✓			6
Paris						✓	✓	✓		✓	✓		5
S2	✓						✓	✓		✓	✓		5
Minos				✓							✓		2
Transib1_moj/wil									✓		✓		2
Uhu						✓						✓	2
Helitron1_Dvir										✓	✓		2
BS2	✓	✓	✓	✓	✓						✓		6
doc	✓	✓	✓	✓									4
doc2	✓	✓	✓	✓	✓								5
doc6	✓	✓	✓	✓									4
FW	✓	✓	✓	✓	✓								5
G6	✓	✓	✓										3
HelenaDS	✓	✓	✓	✓	✓	✓				✓	✓		8
hetA	✓	✓	✓										3
LINE_J1	✓	✓	✓			✓							4
R2	✓	✓	✓	✓		✓							5
TLD1	✓			✓	✓								3
TLD2	✓	✓	✓	✓									4
TLD3	✓	✓	✓										3
1731	✓	✓	✓										3
412	✓	✓	✓										3
Accord	✓	✓	✓										3
Chimpo	✓	✓	✓										3
Quasimodo2	✓	✓	✓										3
Blood	✓	✓	✓										3
Stalker2	✓	✓	✓										3
Tabor	✓	✓	✓										3
Blastopia	✓	✓	✓			✓	✓	✓					6
Copia	✓	✓	✓						✓				4
Gypsy5	✓	✓		✓	✓								4
297	✓	✓	✓	✓	✓	✓							6
Batumi	✓	✓	✓	✓									4
Bel	✓	✓	✓	✓									4
Burdock	✓	✓	✓	✓	✓	✓							6
Circe	✓	✓	✓	✓	✓								5
Diver	✓	✓	✓	✓									4
Diver2	✓	✓	✓	✓									4
Gypsy	✓	✓	✓	✓	✓								5
Gypsy2	✓	✓	✓	✓	✓								5
Gypsy4	✓	✓	✓	✓	✓	✓							6
Gypsy10	✓	✓	✓	✓	✓								5
Gypsy12	✓	✓	✓	✓		✓							5
Invader1	✓	✓	✓	✓	✓	✓							6
Invader4	✓	✓	✓	✓	✓								5
Max	✓	✓	✓	✓									4
Mdg1	✓	✓	✓	✓	✓								5
Nomad	✓	✓	✓	✓	✓	✓							6
Quasimodo	✓				✓								2
Roo	✓	✓	✓	✓	✓	✓	✓	✓					8
Rover	✓	✓	✓	✓	✓								5
Tabor_DA						✓			✓				2
Tirant	✓	✓	✓	✓	✓								5
Transpac	✓	✓	✓	✓	✓	✓							6
Accord2	✓	✓	✓	✓	✓	✓							6
Mdg3	✓	✓	✓	✓									4
TV1										✓	✓		2
Gtwin	✓	✓	✓	✓	✓	✓					✓		7
Gypsy6	✓			✓	✓	✓				✓	✓		6
HMS_Beagle	✓	✓	✓	✓	✓					✓	✓		7
Invader3	✓	✓	✓	✓	✓		✓	✓		✓			8
Invader6	✓	✓	✓	✓							✓		5
Micropia	✓	✓	✓	✓	✓					✓	✓		7
Ninja	✓	✓	✓	✓	✓	✓	✓	✓		✓			9
Osvaldo	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	10
RooA	✓	✓	✓	✓	✓						✓		6
Stalker4	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		10
Zam	✓	✓	✓	✓	✓						✓		6
GypsyDS								✓		✓			2
Total	65	61	60	49	36	24	10	11	4	13	18	2	

Figure 4.22: Distribution of each of the transposable element families for which horizontal transfer was inferred. DNA transposons are shown in blue, non-LTR retrotransposons in purple and LTR retrotransposons in orange.

Figure 4.22 shows the number of transposable element families, for which horizontal transfer is inferred, observed in each species, and in how many species each family occurs. Few cases are detected for *D. grimshawi*, however, as this species is an island endemic, it is assumed it would have fewer opportunities to undergo this process (Clark et al. 2007). This figure clearly illustrates the bias towards *D. melanogaster* and its closest relatives with regards to the current knowledge of transposable elements. The majority of consensus sequences available on Repbase Update correspond to sequences taken from *D. melanogaster*, and as a consequence, these families are often restricted to species of the *melanogaster* group. Although it is possible that these species possess a greater diversity of transposable elements than other members of the *Drosophila* genus, as research has focussed on these species in the past it seems likely that further investigation of the genomes of the species more distantly related to *D. melanogaster* will result in the identification of many new transposable element families. In this study, none of the new families identified were present in the *melanogaster* subgroup.

Generally, families which were present in *D. simulans* exhibited a lower copy number in this species than the others in which they were present, which were most commonly *D. melanogaster* and the sister species of *D. simulans*, *D. sechellia*. The sequences in *D. simulans* also tended to be more degenerate and fragmented. This may be a result of the effective population size of *D. simulans*, which may be large enough for slightly deleterious transposable element integrations to be selected against.

Traditionally, many cases of horizontal transfer were inferred to have occurred on the basis of nothing less than the observation of divergence between

Drosophila host genes exceeding that between transposable elements of the same family in different species (Simmons 1992). The observation of several lines of evidence supporting the hypothesis of horizontal transfer is now taken to provide more conclusive evidence for individual cases (Loreto et al 2008).

There are several limitations associated with the evidence used to infer, and support, the hypothesis that horizontal transfer has occurred. Each of the lines of evidence examined: small divergence, phylogenetic incongruence and patchy distribution, can be attributed to other factors, and therefore do not provide conclusive evidence that horizontal transfer has occurred.

In cases where horizontal transfer has occurred, it is possible that the phylogenies produced using the transposable element sequences are congruent with the host phylogeny. This was observed for 18/74 families, for which horizontal transfer was inferred on the basis of small divergence, but congruent phylogenies were produced. This is particularly the case for transposable element families which are only present in two out of the twelve *Drosophila* species for which the complete genome sequence is available. Where horizontal transfer has occurred and congruent phylogenies are produced, although the topology of the trees is the same, the time to common ancestry of all elements on the tree is shorter than for the host species. Such a phylogeny may be observed where all elements of a particular family currently represented in the genome sequence of the recipient species are descended from the element obtained during the transfer event. Due to the rapid elimination of older elements from *Drosophila* genomes, it is possible that elements that were present before the transfer event have been removed from the genome. Through investigating horizontal transfer of LTR retrotransposons, it was observed that the flanking LTRs of individual elements tend to have very few differences between them, confirming that the majority of elements in the phylogenies constructed are indeed recent integrations. In the case of LTR retrotransposons, the solo LTRs may produce incongruent phylogenies where a congruent phylogeny is observed for the internal portion of the element, where older internal sequences have been deleted but one copy of the divergent flanking LTRs has been retained. This was the case for the Tabor and Stalker2 families, where a congruent phylogeny is produced using internal sequences, but LTR sequences from *D. melanogaster* are found within a clade of sequences from *D. simulans* and *D. sechellia* in both cases. In addition, it is assumed that invading elements have an advantage in that the host may not be equipped to limit their proliferation, such that they proliferate rapidly, while older elements are eliminated from the genome. It is also possible to generate incongruent phylogenies where horizontal transfer has not occurred, for example due to incomplete lineage sorting of an ancestral polymorphism. Assuming that two "subfamilies" of a transposable element family are present in a common ancestor of three species, one of those subfamilies may be lost in each of those three descendent species. If the subfamily which is lost is different between the two most closely related species, this may lead to phylogenetic incongruence.

Examination of phylogenetic incongruence is challenging where small, fragmented and degraded elements are found. It is difficult to confidently align distantly-related elements, particularly where extensive insertions, deletions, truncations and duplications have occurred within the element. Phylogenies suffer from the problem of long branch attraction, and elements belonging to different "subfamilies" may cluster together on a tree of what are, essentially, paralogous elements. In many of the cases discussed, where phylogenetic incongruence is observed, this may be the result of poor alignments of degenerate sequences, which subsequently do not cluster together with other elements from the same species. Therefore, fragmented or highly degraded elements were removed from each sequence set to produce the phylogenies, thus decreasing the likelihood of observing phylogenetic incongruence as an artefact of a factor other than horizontal transfer. Where these phylogenies were incongruent with the host phylogeny, this can be taken as more convincing evidence for horizontal transfer than if degenerate sequences were included. A further limitation of phylogenetic incongruence as evidence for horizontal transfer is the problem of parallel mutation. Particularly where copy numbers are high, parallel mutations may occur within closely-related elements in different species, which could cause them to cluster together on the tree. A further limitation associated with using phylogenetic incongruence as a means of inferring horizontal transfer is the inability of phylogenetic trees to resolve relationships between transposable elements of the same family in very closely-related species, such as *D. simulans* and *D. sechellia*. Particularly where copy numbers are high, elements from these species tend to be jumbled together in a single clade on the phylogeny of the elements. As a consequence, this observation cannot be taken as convincing evidence for horizontal transfer. However, due to the geographical overlap and close relationship between these two species, horizontal transfer is a likely possibility, but is more difficult to confidently support. This limitation may therefore act to reduce the resolution of the analyses to confidently infer horizontal transfer events.

Inferences regarding the direction of putative horizontal transfer events are complicated by the inability to root phylogenetic trees of transposable elements. Unrooted trees may be congruent with the host phylogeny, whereas if the tree could be rooted, this would reveal phylogenetic incongruence. For example, if horizontal transfer has occurred from a donor into a recipient species in which the transposable element family was previously absent, all of the elements in the recipient species would be descended from the transfer event, and would therefore cluster together. This would generate a congruent phylogeny. However, were it possible to root the tree, this clade should fall within the clade of elements from the donor species, and would be indicative of horizontal transfer (Figure 4.23). Due to the relatively long divergence times, it is not possible to root phylogenies of transposable elements using the consensus sequence of a related transposable element family, for example using the sequence of Gtwin to root a tree of Gypsy elements, or Mdg3 to root a tree of Tabor elements. In cases where this was attempted, alignments were poor and ambiguous, and did not provide the information required to reliably root the tree. In addition, the branch leading to the root was extremely long relative to the other branch lengths in the tree, which tend to be very short, connecting elements, which, in many cases, are almost identical. It is also not possible to root the tree using a divergent member of the transposable element family from a species which would form an outgroup to the other species represented on the tree. Without conclusively being able to rule out horizontal transfer involving this species, it could not be assumed that this element would represent an outgroup to the others in the tree.

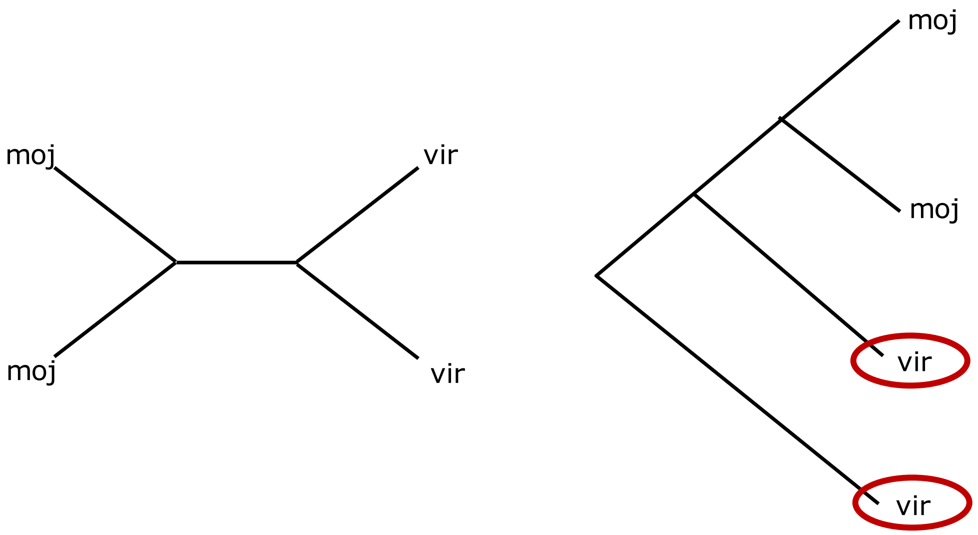


Figure 4.23: Effect of position of the root on the observation of phylogenetic incongruence. The image on the left shows an unrooted tree congruent with known host relationships, When rooted (right), the topology is incongruent with the host phylogeny.

It could be argued that small divergence between transposable elements of the same family in different species may be attributed to constraint. Constraint can operate on transposable elements to the extent that an element which does not encode functional enzymes will not be able to propagate autonomously, and may therefore have a selective disadvantage. However, constraint cannot be observed in the traditional sense, in that there is no selection pressure to maintain a transposable element at a particular locus such that it resembles the element which originally inserted at that locus. However, it can be expected that the majority of elements present would share certain sequence features were these optimal for transposition. As old elements are cleared from the genome rapidly, the recent descendents of transposition events, which must have been functional at the moment of integration and therefore may have certain sequence features in common, may be the only elements observed in the genome for some transposable element families. However, in the putative cases of horizontal transfer described above, small divergence is not restricted to the coding regions of the elements, implying recent common ancestry. For "constraint" to be observed across the full length of the element, there would have to be certain sequences in the noncoding regions, such as enzyme recognition sites or sites involved in RNA secondary structure, which are required for effective transposition. This may indeed be the case, for example the CR1 family of non-LTR retrotransposons possesses sequences in its 3' UTR which may act as a recognition site for reverse transcriptase (Kapitonov and Jurka 2003b), which appears to be constrained between elements. It may also be that, especially where copy numbers of transposable elements are high, few mutations have occurred between elements simply as a result of chance, particularly where the host species are closely related. As discussed previously, the observation of patchy distribution of a particular transposable element family across the *Drosophila* phylogeny is not, alone, convincing evidence for horizontal transfer. This is a result of stochastic loss of entire families from certain lineages as a result of the frequent elimination of transposable elements from *Drosophila* genomes, which can result in entire families being lost. This process appears to have been particularly common for *D. erecta*, from which many transposable element families are absent, which are present in the closely-related species *D. yakuba*, *D. melanogaster*, *D. simulans* and *D. sechellia*.

4.5 Conclusions

An unbiased approach to investigating horizontal transfer in *Drosophila* has revealed that it appears to be a more frequent occurrence than previously anticipated, with good evidence for the process in a total of 74 families out of

141 (52%) which were investigated. Even among the non-LTR retrotransposons, for which horizontal transfer was reported to be very rare due to the mechanism of mobilisation, horizontal transfer appeared to be relatively common, with evidence for the process occurring in 13/41 (32%) families. Three lines of evidence were used to determine whether or not horizontal transfer had occurred. Of these, small divergence was considered to be the most reliable, however, several instances were observed in which divergence was smaller than the host *Adh* genes, but horizontal transfer was unlikely to have occurred, as well as divergence larger than the host genes where horizontal transfer almost certainly had occurred. Phylogenetic incongruence appeared generally to be a reliable method of determining directions of putative transfers, however, was unable to resolve relationships between closely-related species, or transposable elements with little divergence among them. Patchy distribution across the host phylogeny appeared to be much more frequently attributed to stochastic loss of a transposable element family from particular lineages, a common process in *Drosophila*, rather than horizontal transfer. Contrary to previous expectations, the greatest evidence for horizontal transfer was found for the LTR retrotransposons (49/59 families), rather than the DNA transposons (12/41 families), which have previously been reported to undergo the most frequent mobilisation between species (Loreto et al. 2008). This has since been confirmed by the work of Bartolome et al. (2009). The possible explanations for this observation will be discussed in the following chapter.

Chapter 5 - *env* is constrained and may provide a self-encoded vector for horizontal transfer of *Drosophila* LTR retrotransposons

5.1 Introduction

The last chapter discussed the relative frequencies of horizontal transfer among the three main types of transposable element, DNA transposons, LTR retrotransposons and non-LTR retrotransposons, in *Drosophila*. The finding that LTR retrotransposons appear to undergo horizontal transfer more frequently than DNA transposons in *Drosophila*, also reported by Bartolomé et al. (2009) is contrary to estimates based on previous studies, which suggest horizontal transfer is most common for DNA transposons. This was believed to be explained by the relative proportion of the lifecycle that each of these three types spend as a DNA element. DNA transposons exist exclusively as DNA. LTR retrotransposons exist for part of the lifecycle as DNA, following reverse transcription of an RNA copy of the element. This cDNA copy is a free-floating entity, which then integrates into the genome through the integrase activity of the *pol* protein. Although non-LTR retrotransposons also undergo reverse transcription, this occurs at the target site of integration, and is termed target primed reverse transcription (TPRT). Mobilisation via this mechanism results in non-LTR retrotransposons never existing as a free DNA copy.

Under the assumption of vertical transmission of transposable elements with no horizontal transfer, it could be anticipated that divergence among retrotransposons would be higher than for DNA transposons, due to the relatively low fidelity of the reverse transcriptase enzyme, encoded by pol in LTR retrotransposons, and ORF2 in non-LTR retrotransposons. Reverse transcriptase introduces mutations into the sequence of the DNA more frequently than DNA polymerase, the host enzyme that replicates DNA transposons. DNA transposons are replicated during normal cellular DNA replication. Their copy number is increased if an element contained within a region which has already undergone DNA replication is mobilised, and integrates at a new location in a region which has yet to be replicated. Alternatively, the double strand break produced as a result of transposition may be repaired using information from the homologous chromosome, which may reintroduce the element that has been lost. DNA polymerase is a high fidelity enzyme supported by mutation repair mechanisms, which should keep the introduction of mutations to a minimum. This will not be the case for a retrotransposition event. Therefore, it may be expected that divergence between retrotransposons would be higher than between DNA transposons under vertical transmission. This provides further evidence to suggest that horizontal transfer is responsible for the high degree of similarity between LTR retrotransposons of the same family found in different species. However, there are possible explanations for the observation of smaller divergence among the LTR retrotransposons that do not invoke a significantly higher rate of horizontal transfer. These are a greater copy number of LTR retrotransposons relative to DNA transposons, and a greater proportion of individual LTR retrotransposon elements covered by open reading frames. The results presented in this chapter determine that the small divergence between LTR retrotransposons of the same family in different species cannot be attributed to copy number and open reading frame coverage, supporting a higher incidence of horizontal transfer among this group of transposable elements. This high incidence of transfer is not restricted to a particular type of LTR retrotransposons, such as the Gypsy elements. Therefore, the *env* open reading frame is investigated as a possible self-encoded vector, access to which may account for the high rate of horizontal transfer of LTR retrotransposons.

In this chapter, the possibility that possession of an *env* open reading frame may result in a higher frequency of horizontal transfer for the LTR retrotransposons compared with the other two groups is discussed. Alternative explanations for the more frequent observation of small divergence between transposable elements of the same family in different species are investigated, starting with higher copy number of LTR retrotransposons, followed by increased open reading frame coverage, and overrepresentation of Gypsy elements, which may undergo more frequent horizontal transfer compared with other types of LTR retrotransposon. Finally, the possibility of *env* acting as a self-encoded vector is investigated, by examining the env open reading frame for constraint, and the possibility of LTR retrotransposons that do not themselves encode *env* gaining access to the envelope protein produced by another element is discussed.

5.2 Methods

To assess whether copy number has an effect on the smallest divergence observed, firstly, the number of elements from each family in each species was recorded. These values were multiplied together for each interspecies comparison, to give the total number of comparisons. For example, if there were five elements of one family in *D. melanogaster* and six in *D. sechellia*, the total number of comparisons for that family between *D. melanogaster* and *D. sechellia* would be 30. Graphs were then plotted of total comparisons against the smallest divergence observed, for each interspecies comparison. The data were tested for normality using the Kolmogorov-Smirnov test in

SPSS. Where the data were found to be normally distributed, a Pearson correlation was performed in SPSS. Where data were not normally distributed, log transformation was performed to try to produce normally distributed data, to improve the sensitivity of the test and therefore the likelihood of detecting a significant result. Where transformation yielded data that were still non-normal, a Spearman correlation was performed using SPSS. Where a significant correlation was found, and high leverage points appeared to strongly influence the correlation, these were removed and the correlation repeated to test whether the significance could be attributed to a single point. To determine whether or not copy number is indeed higher for LTR retrotransposons compared with the other two groups, mean copy number of elements of each group (LTR retrotransposons, DNA transposons and non-LTR retrotransposons) was calculated for each species using Microsoft Excel. Where the mean copy number of LTR retrotransposons was higher for a particular species compared with either DNA transposons or non-LTR retrotransposons, a one-tailed independent samples t test was used to determine whether the copy number was significantly higher, where data were normally distributed. Else, a one-tailed Mann Whitney U test was performed. A one-tailed test was chosen as the tests were only applied where the copy number of LTR retrotranspons was higher, to determine whether it was significantly so.

To investigate any potential relationship between open reading frame coverage and the amount of divergence observed between transposable elements of the same family in different species, Repbase Reports were used to identify the locations and lengths of open reading frames. Where this information was unavailable for a particular family, NCBI ORF finder was used to identify the location and length of the open reading frames in the consensus sequences for each family. This was used to calculate the percentage of each element that was covered by open reading frames. This was assumed to be the same, or very similar, for all autonomous elements in the family, even if slight variation from the consensus is observed. Although the contemporary elements observed in the genome have generally mutated from this consensus, and therefore may exhibit different open reading frame coverage, it is only the coverage at the moment of integration which is relevant, as following integration the element can be assumed to be unconstrained. Correlations were performed, using the same methods as for copy number, between the percentage open reading frame coverage and smallest divergence observed, for each interspecies comparison.

To look at whether or not *env* is under constraint, the Ka/Ks ratio was calculated using DnaSP (Rozas et al. 2003). Alignments of the *gag, pol* and *env* open reading frames were produced in BioEdit. Nonsense and frameshift mutations cannot be included in the Ka/Ks analysis, therefore any codon affected by a mutation of this nature in any of the sequences in the alignment was removed. These sites were identified by manually scanning the alignments for insertions and deletions that would result in a frameshift. Following removal of any codon affected by a frameshift, sequences were individually entered into the Expasy Translate tool, and translated into amino acid sequence. The location of any stop codons was noted, and these codons were removed from the alignment. This method resulted in the majority of sites being retained for the Ka/Ks analysis, and assuming all sites are equally likely to mutate, should not introduce any bias. Removal of sequences which contained premature stop codons or frameshifts would remove the majority of the sequence data, and is also likely to introduce bias, as these elements are likely to be the most divergent, and only recent insertions would be retained.

Following removal of these codons from the alignment, the Ka and Ks values for each pairwise comparison were calculated using DnaSP. The Ka/Ks ratio was then calculated. Where this value is less than 1, this indicates that the pair of elements under comparison are, or have been, under constraint, as the number of replacement changes that have occurred is fewer than expected given the number of synonymous changes that have occurred. The sequence similarity between members of the Gypsy superfamily was investigated by performing alignment of the consensus sequences for the internal sequence and the LTRs of the 43 Gypsy-like families obtained from Repbase Update, using the ClustalW tool in BioEdit. Identity between the different families was determined by performing pariwise comparisons, with gaps ignored between individual pairs, using MEGA (Kumar et al. 2008). LTR sequences were scanned manually for regions of high identity.

5.3 Results and Discussion

5.3.1 Copy number

Higher copy number could possibly lead to small divergence being observed, as this would increase the probability of two elements, one in each species, having accumulated very few mutations just by chance, assuming that the number of mutations follow a Poisson distribution. For example, assuming the average number of mutations between elements of the same transposable element family in different species follows a Poisson distribution with a mean of 5, at very low copy numbers, the number of mutations between all pairs of elements would be expected to fall around this average. However, if copy number was very high, for example 100 elements in each species, it would be expected, under the Poisson distribution, that numbers of mutations of only 1 or even 0 would be observed just by chance. 0 mutations would indicate identical elements, which would suggest horizontal transfer had occurred. If two species A and B are compared, and there was one element per species, and five mutations are expected to have occurred, the Poisson probability of having zero mutations would be 0.0067. However, if the two species each have *n* elements, and if the *n* elements in each species have a star-shaped tree and a time to most recent common ancestor that is half way to the time to common ancestry of elements from A and B, the probability that there will exist a pair of elements, one from A and one from B, with zero mutations separating them, is given by $e^{-2.5}$ x $(1-(1-e^{-1.25})^{n})^{2}$ (Figure 5.1). With *n*=100 , this probability is 0.0821, with *n*=10, it is 0.0766, with *n*=2 it is 0.0198 , and with *n*=1, it is 0.0067, as expected.

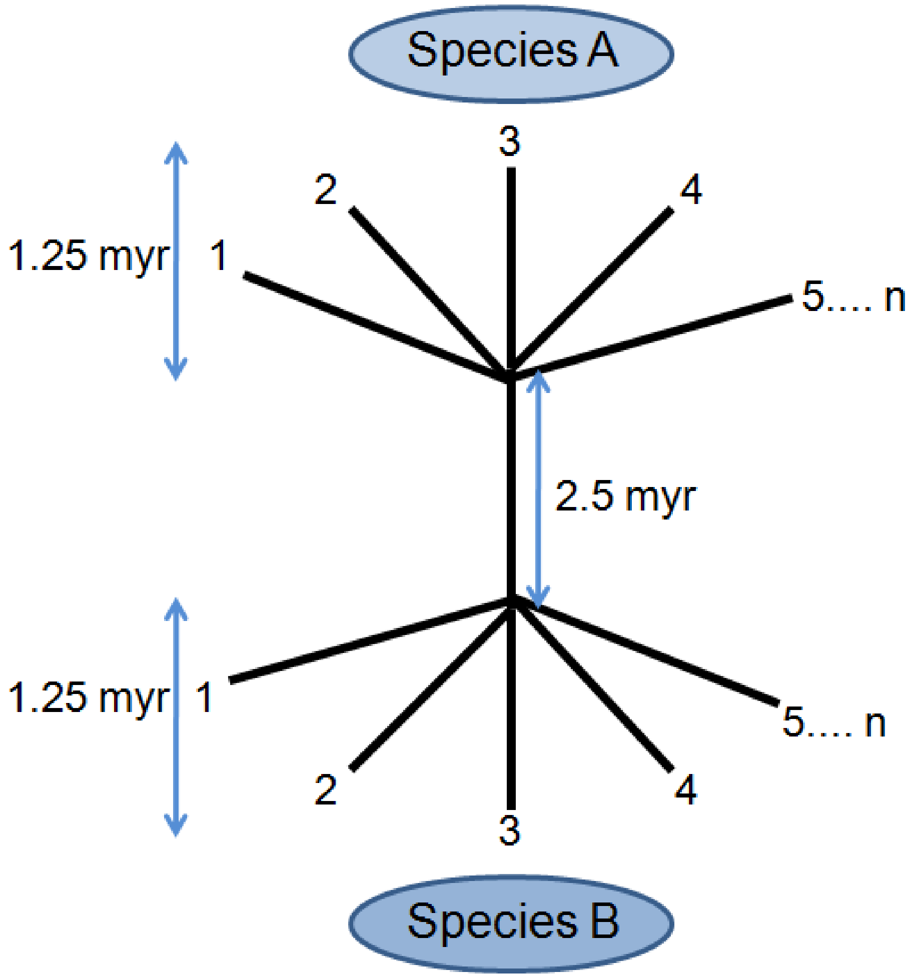


Figure 5.1: The probability of two elements from the same transposable elements in two different species, A and B, sharing zero mutations, given a star-like tree, divergence time between species A and B of five million years, and the time to most recent common ancestry of two elements from A and B of 2.5 million years. Half of the evolution occurs on the central branch and a quarter on each of the radiations in species A and B. Given a Poisson distribution, if five mutations are expected, the probability of zero mutations occurring on the central branch, which represents half of the evolution (i.e. half of the expected mutations should occur on this branch) is given by $e^{-2.5}$. The probability that any one of the terminal branches has zero mutations is $e^{-1.25}$, therefore the probability of having at least one mutation is $1-e^{-1.25}$. The probability of all one of the terminal branches in one species having at least one mutation is given by $(1-e^{-1.25})^n$. Consequently, the probability that at least one element does not have at least one mutation, i.e. at least one element has zero mutations, is given by $1-(1-e^{-1.25})^n$. This is squared as this must be true in both species A and B.

This effect is indeed observed (Figure 5.2), as there is a significant negative correlation between the total number of comparisons of LTR retrotransposons of the same family between two species, and the smallest amount of divergence detected (e.g. *D. melanogaster-D. simulans*, r_s = -0.401, p = 0.002). However, mean copy number of LTR retrotransposon families in each species is often lower than for DNA transposons and non-LTR retrotransposons. In cases where it is higher, it is not significantly so (e.g. *D. sechellia*, 18.3 compared with 17.8, p= 0.975).

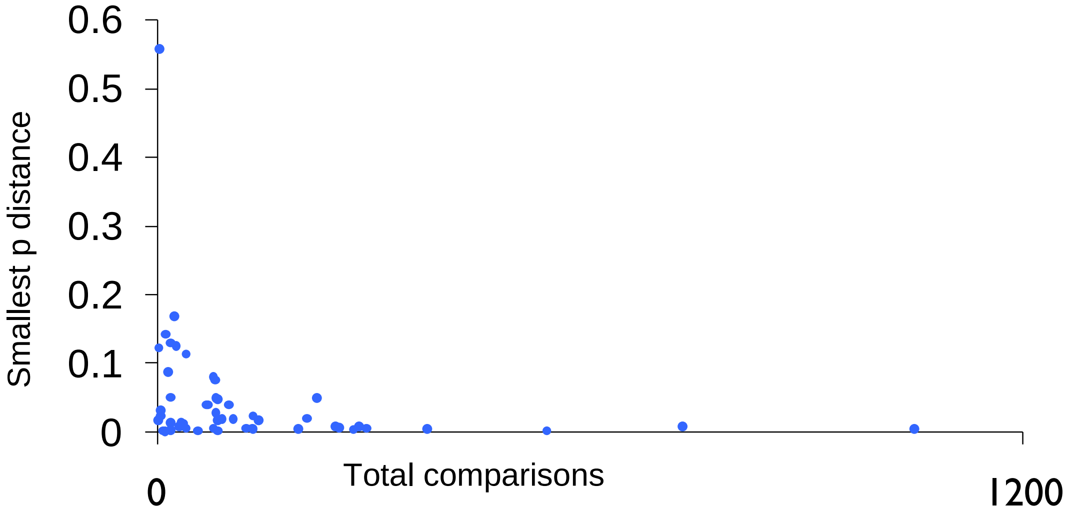


Figure 5.2: Correlation between the total number of comparisons between species for each LTR retrotransposon family, and the smallest divergence observed between *D. melanogaster* and *D. simulans*. A significant negative correlation is observed, such that as the copy number, and therefore number of comparisons, increases, the smallest divergence observed decreases. Removal of the single, apparently high leverage, point, has a negligible effect on the correlation and renders it slightly more significantly negative.

5.3.2 Open reading frame coverage

Another possibility is that open reading frames, on average, cover a greater proportion of an LTR retrotransposon element than a DNA transposon. As these open reading frames can be assumed to be under constraint up until the moment of integration, at least in the case of *gag* and *pol,* this may reduce the amount of divergence between these elements in different species, compared with DNA transposons. There is a significant negative correlation (e.g. *D. melanogaster-D. simulans* divergence, $r_s = -0.438$, $p < 0.001$) between the proportion of the family consensus sequence covered by open reading frames and the smallest divergence seen between two elements of the same family in different species (Figure 5.3).

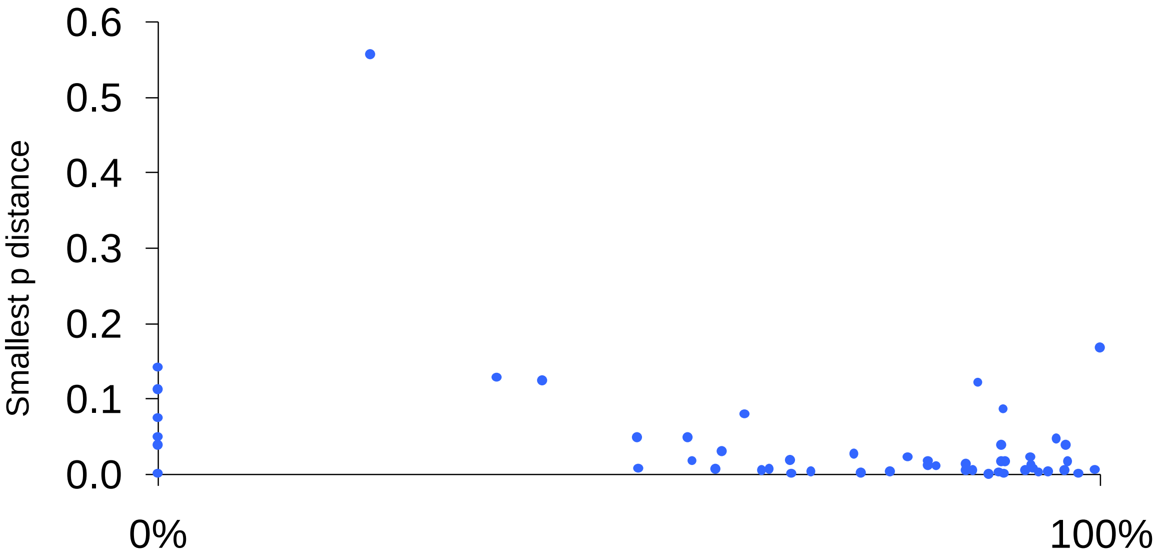


Figure 5.3: The correlation between percentage ORF coverage and smallest divergence observed for LTR retrotransposons compared between *D. melanogaster* and *D. simulans*. When the single, apparently high leverage, point is removed, the correlation is unaffected and the p value increases slightly, rendering the correlation more significant.

Therefore, as the extent of the open reading frame coverage increases, the smallest amount of divergence seen decreases. This may have an effect on the observation that smaller divergence is observed between LTR retrotransposons in different species, as these elements have a total open reading frame coverage of 70.1%, i.e. this percentage of the concatenated length of the LTR retrotransposons included in this study is covered by open reading frames. This is significantly higher ($p = 0.0405$) than the 57.8% total open reading frame coverage of DNA transposons. However, if only LTR retrotransposons which have an open reading frame coverage less than 57.8% are considered, a greater proportion of families (69%) showing small divergence between species is still observed than for DNA transposons (29%). Furthermore, the open reading frame coverage for LTR retrotransposons includes *env,* which is not essential for retrotransposition within a cell, as discussed below. Exclusion of *env* yields a total open reading frame coverage of 64.1% for LTR retrotransposons, which is not significantly higher than for DNA transposons ($p > 0.05$). The total open reading frame coverage of non-LTR retrotransposons, 68.0%, is not significantly different from that of the LTR retrotransposons, including or excluding *env.* Therefore, the more extensive open reading frame coverage observed for LTR retrotransposons cannot account for the frequency with which divergence smaller than between the host *Adh* coding regions is observed for this type of transposable element in *Drosophila*. It does, however, appear that the extent of the open reading frames has an effect on the level of divergence, and the extensive coverage in LTR retrotransposons may inflate the estimate of horizontal transfer for these types of elements.

Type of transposable element	Total ORF coverage (%)	Average ORF coverage (%)
LTR retrotransposon (including env)	70.1	69.7
LTR retrotransposon (excluding env)	64.1	64.5
Non-LTR retrotransposon	68.0	65.2
DNA transposon	57.8	47.6

Table 5.1: Open reading frame coverage of the three types of transposable element. Total open reading frame coverage corresponds to the total open reading frame length as a fraction of the total length, whereas the average open reading frame coverage corresponds to the average of the percentage coverage of individual families.

5.3.3 Overrepresentation of Gypsy

Gypsy-like elements are overrepresented in the LTR retrotransposon dataset, with 43 of the 59 families investigated belonging to the Gypsy superfamily. It is possible that in reality, only Gypsy-like elements undergo frequent horizontal transfer among the LTR retrotransposons, and their overrepresentation would result in the appearance of horizontal transfer being a common occurrence for the entire group. A greater proportion of Gypsy-like families (86%) do appear to have been involved in at least one horizontal transfer event when compared with LTR retrotransposons belonging to other superfamilies. However, of the LTR retrotransposons which belong to other superfamilies, 75% appear to have been involved in at least one horizontal transfer event. This is still much higher than the proportion of DNA transposons (29%). Therefore, it is not the case that, among the LTR retrotransposons, only Gypsy-like elements are more likely to have undergone at least one horizontal transfer event.

5.3.4 *Env* as a self-encoded vector

Therefore, it appears that the observation that LTR retrotransposons tend to have divergence smaller than that of *Drosophila* host genes, is most likely explained by horizontal transfer, which is frequent among all superfamilies. A possible explanation that has been discussed for this observation is some LTR retrotransposons have retained the capacity to encode *env,* and therefore produce a virus-like particle (VLP), that could be used as a self-encoded vector to transfer the element to another species (Llorens et al. 2008). 21 out of the 59 LTR retrotransposon families encode *env,* either as a separate open reading frame, or, in the case of the Roo family, as part of a single *gag-pol-env* polyprotein. To investigate the possibility that *env* is providing a vector for horizontal transfer, the level of constraint in the *env* oepn reading frame was investigated. Constraint in *env* would indicate that it is being selectively maintained to perform an advantageous function. However, if *env* performed a function involved in retrotransposition within a cell, this would also lead to constraint, and therefore constraint in *env* would provide no evidence to suggest involvement in horizontal transfer. It is unlikely that *env* is advantageous for retrotransposition within a cell, due to the fact that 64% of the LTR retrotransposon families investigated do not encode *env.* 98% have retained *gag* and *pol,* and therefore the ability to propagate autonomously.

To investigate the possibility that *env* is involved in intracellular retrotransposition, *env* open reading frames were examined for the presence of shared premature stop codons and frameshift mutations that would render the *env* protein non-functional. If groups of these mutations are shared among different elements of the same LTR retrotransposon family in the same species, this would indicate their propagation by retrotransposition rather than parallel mutation, and therefore that *env* is not required for successful retrotransposition within a cell. Six of the 21 families which encode *env* were not included in the analysis. Roo was excluded as *env* is encoded as part of a single *gag-pol-env* polyprotein, and therefore constraint on *env* would be difficult to elucidate. Although the region which results in *env*-like activity in the polyprotein could be isolated, this region may be constrained due to effects on, for example, protein folding, which would affect the *gag*-like and *pol*-like activity of the protein as well. Families 176, Gypsy3, and Tom were excluded as these families are only present in one of the twelve *Drosophila* species for which the sequenced genome is available. Therefore there are no sequences to use as a comparison to determine whether or not the open reading frames are under constraint. Clearly, as these families are only present in a single species, there is no evidence for horizontal transfer for these families. TV1 was excluded from the analysis as, although it is present in two species, the *env* open reading frame is only intact in *D. virilis*. Finally, Gypsy9 was excluded as only remnants of its open reading frames are intact, and none of the elements appear to be capable of retrotransposing autonomously. It is possible that Gypsy9 no longer actively mobilises, which is supported by the divergence of the flanking LTRs in Gypsy9 elements. According to the Repbase Report for Gypsy9 (Kapitonov and Jurka 2002a), the elements in *D. melanogaster* are flanked by 6% divergent LTRs, therefore it is assumed that the family stopped proliferating in this species around four million years ago. Therefore, it does appear that this family has been vertically transmitted and is no longer proliferating, which is consistent with the divergence data.

The lack of a requirement for functional *env* for intracellular retrotransposition is suggested both by the absence of *env* from a large number of successful LTR retrotransposon families, and also the presence of shared stop codon and frameshift mutations in the *env* open reading frame. These mutations are likely to have been present in an element prior to its retrotransposition, and are subsequently observed in the daughter element. As mutations of this nature would render *env* non-functional, this suggests *env* is not required for intracellular retrotransposition. However, *gag, pol* and *gag-pol* open reading frames were also examined for the presence of shared nonsense and frameshift mutations, and in several instances such mutations were observed. Examination of flanking DNA sequences reveals that, for example, in the case of Idefix *gag-pol,* a premature stop codon shared by two elements in *D. sechellia* appears to be attributed to replication of a single element during a larger duplication event. However, in other instances, such as Rover *gag-pol,* a shared single base insertion cannot be attributed to a duplication event. This could be explained by parallel mutation, particularly as, as in most cases, the mutation is shared by only two elements. In some cases, such as Osvaldo, the elements are quite divergent from each other, making parallel mutation a plausible explanation for the shared mutations in *gag* and *pol.* However, it is possible that these elements have been able to retrotranspose in the absence of functional *gag, pol,* or *gag-pol,* which therefore might also apply to *env,* in that LTR retrotransposons may be able to utilise envelope proteins produced by other elements.

Family	ORF	Species	Elements	Shared mutation(s)	Duplicates?
Gtwin	*env*	*D. melanogaster*	2	STOP	No
		D. erecta	2	1bp del	No
Idefix	*gag-pol*	*D. sechellia*	2	STOP	Yes
Gypsy4	*env*	*D. melanogaster*	2	7bp del	Yes
		D. sechellia	2	8bp del	No
Rover	*gag-pol*	*D. melanogaster*	2	1bp ins	No
Tirant	*gag*	*D. sechellia*	2	10bp del, STOP	No
Gypsy	*pol*	*D. sechellia*	2	46bp del, 114bp del	No
297	*env*	*D. melanogaster*	2	STOP, 17bp del	No
	gag	*D. simulans*	2	1bp del	No
		D. sechellia	3	1bp ins	No
Osvaldo	*gag*	*D. simulans*	2	22bp del	No
		D. sechellia	2	80bp del	No
	pol	*D. sechellia*	3	STOP	No
Quasimodo	*env*	*D. melanogaster*	2	1bp ins, STOP	No

Table 5.2: Shared nonsense and frameshift mutations in LTR retrotransposon families possessing *env.* The number of elements which share the mutation, and the type of mutation, are given. Whether or not the shared mutations can be attributed to a duplication, rather than retrotransposition, event is shown.

Analysis of the Ka/Ks ratio in *env* reveals that *env* is under constraint, with the average Ka/Ks for the *env* open reading frame for each family calculated to be less than 1 (Table 5.3). Within each family, there is variation in the values of Ka/Ks, with some comparisons yielding Ka/Ks around 1, suggesting *env* is not constrained in these elements. In many cases, this was true of pairs of elements which shared high identity, with large values of Ka/Ks brought about by extremely small values for both Ka and Ks. For example, out of the 125 pairwise comparisons of the *env* gene of the 297 family, 34 yield Ka/Ks greater than or equal to 1, but in all cases, this can be attributed to very small values of both Ka and Ks. In sixteen of these comparisons, Ks is actually 0, indicating no synonymous changes and extremely recent common ancestry. It appears that due to chance, these elements have accumulated a very small number of nonsynonymous rather than synonymous changes in the short time since their divergence. It can be assumed that, over time, synonymous changes will accumulate. There are some pairwise comparisons between elements of the Gypsy4 and Rover families for which Ka/Ks is around 1, which may be attributed to lack of constraint on the *env* open reading frames. For example, in a pairwise comparison between a Gypsy4 element in *D. melanogaster* and one in *D. sechellia*, Ks and Ka are equal to 0.0223 and 0.0224, respectively, yielding Ka/Ks of 1.0045. Comparison of a Rover element from *D. yakuba* with one from *D. erecta* yields Ka/Ks of 1.107, from Ka of 0.0424 and Ks of 0.0383. There are many other similar cases for both of these families, all of which may be attributed to lack of constraint between the elements.

The mean family Ka/Ks ratio for *env* is 0.2687 and the mean Ka/Ks for all LTR retrotransposon elements extracted that encode *env* is 0.2839. This suggests that *env* does indeed have a selectively-maintained advantageous function. A logical function for *env,* given its role in retroviruses, would be the provision of a self-encoded vector for horizontal transfer. As transposable elements are readily eliminated from *Drosophila* genomes, there may be strong selection pressure to infect another species to ensure continued survival. Constraint on the *env* open reading frame supports the suggestion that horizontal transfer is an essential part of the lifecycle of many transposable elements in *Drosophila*. Access to a self-encoded vector would increase the frequency of horizontal transfer, possibly allowing elements to infect naïve genomes, which cannot control their proliferation. Over time, these species will evolve appropriate defence mechanisms to the invading transposable element, but during this time the transposons would be free to proliferate.

Family	**Mean Ka/Ks**			
	env	***gag***	***pol***	***gag-pol***
297	0.6288	0.85411	0.37478	
Gtwin	0.1597	0.2582	0.1349	
Gypsy	0.1277	0.15624	0.08266	
GypsyDS	0.13494	0.14021	0.10969	
Gypsy4	0.84896	0.48189	0.69584	
Gypsy5	0.23344	0.57967	0.28679	
Gypsy6	0.26798	0.40836	0.22727	
Gypsy10	0.19919	0.3058	0.1521	
Idefix	0.13327			0.06393
Osvaldo	0.06785	0.18591	0.13217	
Quasimodo	0.34697	0.24797	0.18503	
Rover	0.18466			0.13328
Tirant	0.16584	0.19837	0.18296	
Zam	0.2622			0.18055
Average	**0.2687**	**0.3470**	**0.2331**	**0.1259**

Table 5.3: Mean Ka/Ks values for the four open reading frames that can be present in LTR retrotransposons, shown for each of the fourteen families for which *env* is present, and Ka/Ks is calculable. The mean is taken of all Ka/Ks values for each family, for each possible individual interspecies comparison, across all species in which the family is present. The data used to calculate these averages can be found in Supplementary Data 3. Accord is excluded as in the vast majority of pairwise comparisons, Ks = 0, such that Ka/Ks cannot be calculated. Gypsy9 is excluded as intact open reading frames are absent from all elements. The remaining five families which encode *env* but are not included in the above table are only found in a single species, or contain *env* within a single *gag-pol-env* open reading frame.

In investigation of pairs of recent LTR retrotransposon integrations in two species, where the flanking LTRs are 100% identical to each other in both elements of the pair, it is expected that a greater variance in Ka/Ks would be observed for *env* than for *gag* and *pol.* This is because *gag* and pol are constrained until the moment of integration, and are then free to evolve under no selective constraint. In recent integrations, this period of neutral evolution would be short, as *gag* and *pol* must have been functional at the moment of integration. *Env*, however, may have lost function and begun evolving neutrally at any point prior to the integration event. For example, it may have provided a vector for a relatively ancient horizontal transfer event, and have been evolving neutrally ever since, and would therefore be expected to have Ka/Ks closer to 1. Alternatively, *env* may have lost function very recently, if it were involved in a horizontal transfer event occurring just prior to the recent integrations. However, the variance in *env* Ka/Ks is not significantly higher than the variance in *pol* Ka/Ks (Levene's test, p = 0.629). This is in spite of the fact that the variance in *env* Ka/Ks may be expected to be higher than *pol* due to its shorter sequence length relative to *pol,* as small differences in the number of mutations would generate large differences in Ka/Ks variance for short sequences.

The variance in *gag* is significantly smaller than both *env* and *pol.* This is contrary to the expectation that the variance would be similar to *pol* as the two open reading frames have been able to evolve neutrally for the same amount of time. Additionally, as with *env,* the sequence of the *gag* open reading frame is smaller than that of *pol,* which could potentially lead to an increase, rather than a decrease, in the variance in Ka/Ks. A likely explanation is that a greater proportion of *gag* amino acids are unable to mutate without rendering *gag* non-functional. This would result in *gag* being under greater constraint than *pol,* and therefore *pol* may have acquired a larger value of Ka/Ks prior to the integration event, due to a greater proportion of amino acid replacement changes, which increase the value of Ka. This would increase the variability among the values of Ka/Ks for *pol* relative to *gag.* It has to be borne in mind that the number of replacement changes able to occur in *env* without eliminating function, compared to the number able to occur in *pol,* is unknown. However, a greater variance would still be expected for *env* even if it is under greater constraint than *pol,* due to the potentially huge variation in the time since it was last involved in a horizontal transfer event.

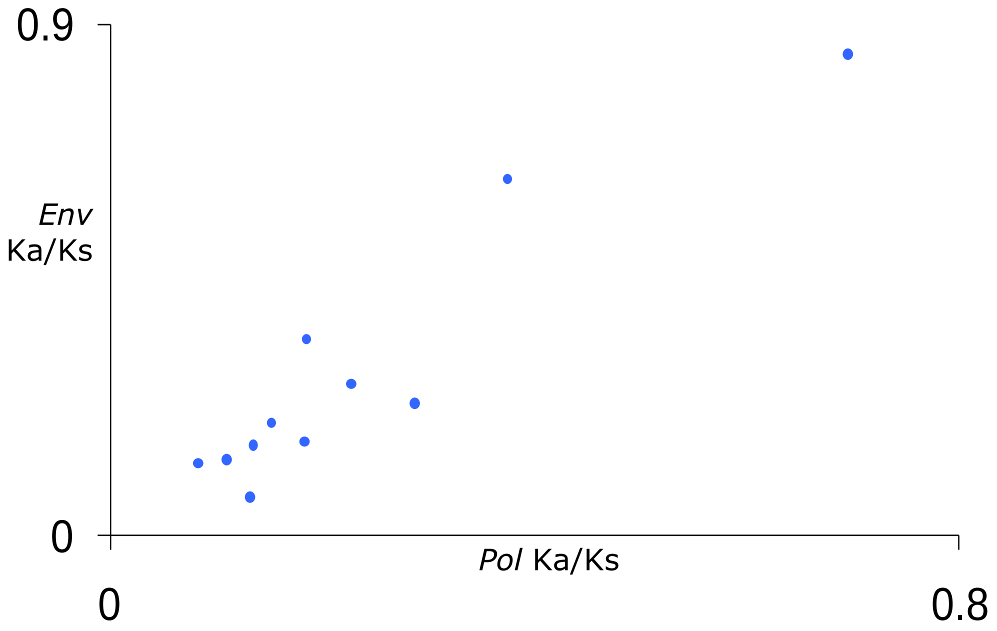


Figure 5.4: Correlation across families of the Ka/Ks values for the *pol* and *env* open reading frames.

A strong significant correlation is observed between *env* and *pol* Ka/Ks (r = 0.871, p = 0.001, Figure 5.4). Consequently, the value of Ka/Ks for *env* can be closely approximated simply through knowledge of the Ka/Ks value for *pol*. Additionally, significant correlations are observed between *gag* and *pol* Ka/Ks (r = 0.839, p = 0.002), and *gag* and *env* Ka/Ks (r = 0.763, p = 0.01). This may suggest that the recent integrations included in this study are descended from recent horizontal transfer events, and, as such, *env* did not lose function until shortly before *pol,* that is, at the time of the integration event. This might suggest that horizontal transfer is an even more frequent occurrence than previously expected. If recent integrations are the product of recent transfers, this suggests that retrotranspositions of elements that have been introduced into the genome by relatively ancient transfers are rare, such that the current LTR retrotransposon residents of the genome are new arrivals, and other elements have been deleted or have mutated to the extent that they are not capable of autonomous proliferation. It also suggests, in keeping with the high level of deletion of transposable elements in *Drosophila*, that elements do not tend to transfer into a genome and remain inactive for an extended period of time. It is likely that any element that successfully transferred but did not proliferate rapidly would be lost, and therefore would not contribute to the current transposable element population in the genome.

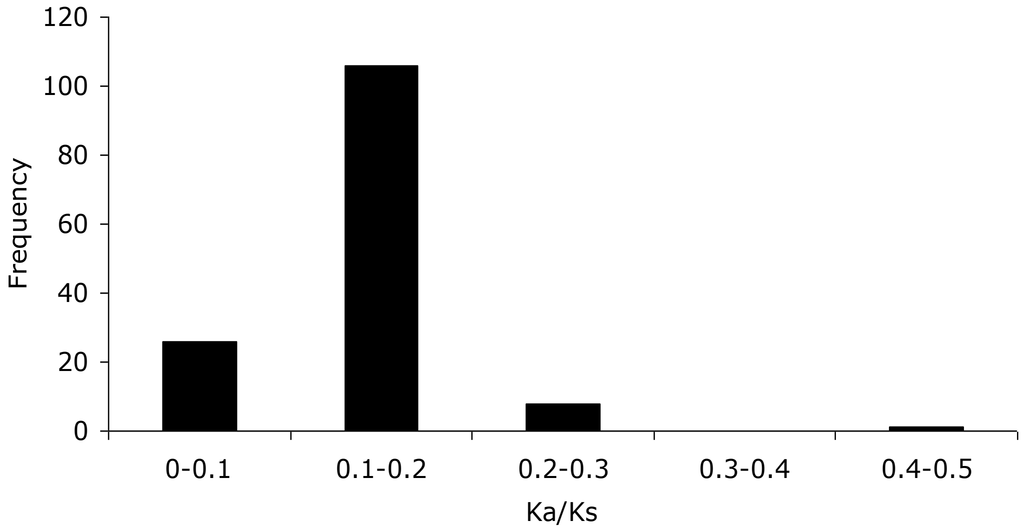


Figure 5.5: This chart shows the distribution of Ka/Ks values among interspecies comparisons for *env* for the Idefix family. Although there is a large range in values, from 0.092 to 0.408, the majority of pairwise comparisons yield Ka/Ks values less than 0.2.

Env was found to be constrained in the LTR retrotransposon family Idefix, for which there is currently no evidence for horizontal transfer. Mean Ka/Ks for Idefix *env* is 0.134, with a range of 0.092 to 0.408 (Figure 5.5). If the observation of constrained *env* among the majority of elements in a family can be taken as evidence for recent horizontal transfer, these families can be assumed to have undergone recent transfer, which therefore further increases the proportion of LTR retrotransposon families for which there is evidence for this process. In these cases, transfer can be inferred to have occurred from an unknown donor species. Transfer of Idefix from the species in which it is known to be present cannot be reliably inferred. However, as *env* is functional in Idefix elements in these species, it is likely that horizontal transfers to unknown recipients have also occurred. Constraint in the *env* open reading frame between a limited number of pairs of elements of a particular species for which there is no evidence for horizontal transfer would not provide such strong evidence as in the case of Idefix, where all pairwise comparisons demonstrate constraint in *env* (Figure 5.5). Maintenance of function in a limited subset of elements could simply be due to chance. However, if the frequency of horizontal transfer is indeed very high, a single element possessing functional *env* may be enough to infer horizontal transfer, such that if *env* is functional, the element will infect other species. Many of these events may not be detected, however, due to stochastic loss from the recipient species or transfer within the same species. Such intraspecific transfer might still lead to constraint in the *env* gene, as there is considerable polymorphism between members of the same species in terms of their transposable element content. It may be that horizontal transfer within a species provides a means for a transposable element family to remain established in that species.

Horizontal transfer of families for which *env* is constrained but no evidence has been obtained is a reasonable expectation, as only twelve *Drosophila* genome sequences are available, and there are consequently thousands of other species, for which genome sequences are not present in the sequence databases, which are potential donors and recipients of transposable elements. As the amount of genomic data available for the *Drosophila* genus increases, it is likely that estimates of the frequency of horizontal transfer, which was once considered a rare event, will also increase. Given that 83% of LTR retrotransposons appear to have undergone horizontal transfer within the twelve species for which complete genome sequences are available, it is possible that all families which are capable of autonomous proliferation will be found to have undergone horizontal transfer as further sequence data become available.

Only 21 of the 59 (36%) LTR retrotransposons included in this study encode *env.* It has been argued that as horizontal transfer is as common for LTR retrotransposons that do not encode *env* as for those that do, that this cannot explain the observation that horizontal transfer is most common for the LTR retrotransposons as a group (Bartolome et al. 2009). In fact, the proportion of families which do not encode *env* for which there is evidence for horizontal transfer is slightly higher than for those that encode *env* (33/38, or 87% compared with 16/21, or 76%). As *env* is under constraint, and is not required for retrotransposition within a cell, it is likely that *env* is being used to provide a vector for horizontal transfer. It is possible that elements which do not themselves encode *env* are able to gain access to virus-like particles (VLPs) produced by related elements. For example, it has been suggested that Tabor, which does not encode the *env* protein, might be infectious through interaction with Gypsy or Gypsy-like elements which do encode functional envelope protein (Jurka et al. 2005). Gypsy-like elements, including Tabor, tend to have high sequence similarity, but only 20 of the 43 (46.5%) Gypsy-like families encode *env.* These account for 20 of the 21 LTR retrotransposon families in total which encode *env,* with the remaining family, Roo, belonging to the Bel superfamily. It is possible that the remaining 23 Gypsy-like families which do not encode *env* are able to access the *env* of the 20 that do.

The similarity between the consensus sequences of the Gypsy-like families which were included in the investigation was deduced. Pairwise comparisons of the consensus sequences of Gypsy-like families indicate that many of these sequences, a total of 25 comparisons, share greater than 60% identity. The most closely related families are Stalker2 and Stalker4, which share 77.5% identity. There is a large group, the Tabor group, among the Gypsy-like elements, the members of which share high identity. This group includes Stalker2 and Stalker4, along with Tabor, 412, Blood, Tabor_DA, Mdg1 and Mdg3. None of these families encode *env,* and therefore would not be able to gain access to *env* from a very closely related family. The pairwise comparisons show high identity between only two Gypsy-like families, Accord and Accord2, of which one encodes *env* and the other does not. These families share 65% identity, and it is possible that Accord2 has accessed the envelope protein encoded by Accord in order to infect other species. Therefore, if envelope proteins produced by Gypsy-like elements have indeed been generally accessed by other Gypsy-like elements, this would require recognition of the elements despite relatively large divergence of more than 40%. It is possible that it is only certain regions of the element that are required for recognition, and these may share greater identity, or consistently contain key residues for recognition. It may be that high identity is only required at, for example, the very start or end of an element to be recognised by a related envelope protein, therefore identity between the LTR consensus sequences of the Gypsy-like families was investigated. There is a great deal of similarity between the first 14 base pairs of the LTR. 23 of the 43 Gypsy-like LTR sequences begin with the same five residues, fifteen starting with AGTTA, and a further eight starting with TGTAG. fourteen of the twenty remaining families possess LTRs of which the first 5bp vary at only one position from one of these two common sequences. The first 11bp of the LTRs of Quasimodo2 and 297 are identical to each other, and these two LTRs also share the first 10bp with the LTR of TV1. Both TV1 and 297 encode *env,* but Quasimodo2 does not, therefore if the recognition site for *env* falls within this region, it is possible that Quasimodo2 may have horizontally transferred using the envelope protein of TV1 or 297 as a vector. TV1 also shares 100% identity with the first 13bp of Idefix. 10 of the first 14bp of the LTR are identical between Invader5 and Mdg1, and further similarity is observed between Gypsy12 and Invader2, which are identical across the first 10bp with the exception of a single base deletion in Gypsy12. 11 of the first 14bp of the LTR are identical between Accord and Accord2, however, as discussed above, these families share relatively high identity across their entire length. Osvaldo, which encodes *env,* shares 10 of the first 14bp of the LTR with Invader3, which does not. Further examples of very high identity are observed between families which either both encode *env,* or both do not encode *env,* such as Gypsy and Gtwin, which are identical across the first 13bp, and Blood and 412, and Gypsy7 and Burdock, which share 13 out of the first 14bp of the LTR. The high identity shared by these Gypsy-like elements in this region at the beginning of the LTR may suggest a more essential role in intracellular retrotransposition, such as recognition by the *gag* and *pol* proteins, in addition to potential recognition by the envelope protein.

The ability of LTR retrotransposon families which do not themselves encode *env* to access the envelope protein of related elements may enable these elements to transfer at a higher rate than other classes of transposable element, without themselves producing a self-encoded vector. It would be expected that the elements which do encode *env* would evolve a form of cis-preference of *env* for the DNA encoded by the same element, however, this is not supported by the observation that elements which do not encode *env* do not appear to transfer less frequently. It is also expected that there would be selection pressure on elements which do not encode *env* to be able to overcome any cis-preference of the virus-like particle proteins to gain successful access. Further investigation into how the virus-like particle recognises the genetic material to be packaged, along with possible constraint at these recognition sites in elements that do not encode *env,* could potentially resolve the issue of whether or not virus-like particles are being appropriated by related elements which do not themselves possess an open reading frame for *env.*

5.4 Conclusions

It is possible that the ability of many LTR retrotransposon families to encode *env* is responsible for the higher rate of horizontal transfer observed for this type of transposable element, as other possibilities, i.e. greater open reading frame coverage and higher copy numbers, have been shown not to account for the small divergence observed between members of the same LTR retrotransposon family in different species. *Env* is constrained, despite a lack of a requirement for the envelope protein in intracellular retrotransposition, and therefore may be providing a self-encoded vector, enabling LTR retrotransposons to undergo horizontal transfer at a higher rate than either non-LTR retrotransposons or DNA transposons. It is possible that those LTR retrotransposons that do not themselves encode *env* are able to gain access to the envelope protein produced by elements belonging to related families.

Chapter 6 - Discussion

Through utilising genomic sequence data, the evolution of transposable elements in two contrasting systems, humans and *Drosophila*, has been investigated. In the human system, where retention of transposable elements is commonplace and deletion is a relatively rare event, entire subfamilies of transposable elements can be assembled. This allows models of the amplification of those subfamilies to be constructed to make inferences about their evolutionary history, such as how many elements in each subfamily are capable of transposition. Analysis of several young Alu subfamilies revealed that estimates of the number of elements capable of acting as source elements varies considerably between different subfamilies, and also between species where a subfamily is shared between humans and chimpanzees. In *Drosophila*, where there is rapid turnover of transposable elements, it is not possible to obtain complete transposable element families, nor make confident inferences about the relationships between individual elements within a family through examination of sequence data. Due to the rapid elimination of transposable elements from the genome, horizontal transfer of elements from a donor species into a recipient species is relatively commonplace, and may provide a significant means of escape and survival for these elements. This does indeed appear to be the case, as it was found that horizontal transfer among the *Drosophila* species is a frequent occurrence for all types of transposable element. Horizontal transfer was found to be most common for the LTR retrotransposons, which may be attributed to the ability of some elements of this type to produce virus-like particles through activity of the *env* gene, which was shown to be constrained.

The factors affecting the evolution of transposable elements depend greatly on the genomic environment in which those elements reside. The transposable element composition of humans and *Drosophila* is similar in that representatives of both classes of transposable element are present and actively transposing, with some, such as the Mariner-like elements, common to both species. However, the proportion of families present which are currently active differs between the species, with, for example, L1 the only active autonomous non-LTR retrotransposon in humans. In *Drosophila*, the vast majority of families appear to be active, and show relatively low divergence amongst elements of the same family, probably as a consequence of the strong selection pressure on, and rapid turnover of, elements. In humans, copy numbers of individual transposable element families tend to be greater than in *Drosophila*, due to the general retention of elements, such that a more limited number of families was investigated. As this population of transposable elements is relatively stable, comparisons can be drawn between closely-related species such as humans and chimpanzees, allowing conclusions to be drawn such as an increase in retrotransposition rate along the human lineage, along with the identification of complete gene conversion events and recent integrations. In *Drosophila*, horizontal transfer of elements belonging to a large proportion of families can be assumed to have occurred, whereas in humans this process is not part of the transposable element lifecycle. This may be due to lack of opportunity for horizontal transfer, perhaps the lifecycle and behaviour of *Drosophila* are more conducive to the process, for example through hybridisation of different species. Alternatively, lack of horizontal transfer of transposable elements in humans could be due to a lack of a requirement for the process by the elements themselves, as they are not as strongly selected against, and tend to survive in the genome. In *Drosophila*, without horizontal transfer, the vast majority of transposable element families would have been eliminated, as evidenced by the presence of only recent insertions of many families in many of the genomes, and by the recent stochastic loss of families from some lineages. Furthermore, it is possible that there are more effective mechanisms in place in the human genome to resist the invasion of transposable elements introduced horizontally, or that transposable elements in *Drosophila* have overcome such measures as a consequence of the selection pressure on them to be able to infect other species.

Over evolutionary time, individual transposable element insertions become either fixed in a population or lost by genetic drift. During the period between integration and either fixation or elimination, the particular transposable element insertion is polymorphic for presence or absence within the population. In humans, the insertion polymorphism level of Alu elements provided evidence to support the activity of secondary source elements. Where mutations are shared among polymorphic elements, this suggests these mutations are present in active source elements and have been propagated by retrotransposition rather than parallel mutation or gene conversion. The presence of polymorphic elements also confirmed that a young subfamily is still actively retrotransposing, and has been used in other studies to investigate population structure. However, although polymorphism is a useful observation in examining transposable element evolution in humans, the vast majority of insertions are fixed in the population, and are shared between individuals. In *Drosophila*, the vast majority of transposable element insertions are not fixed between individuals. Due to the rapid turnover of elements, many elements are either in the process of being lost from the population, or are recent integrations possessed by only a few individuals. When comparing two members of the same species, such as *D.*

melanogaster, it is unexpected that the genome sequences of two individuals would have the same transposable element composition. Furthermore, orthologous elements are rarely detected between closely-related species such as *D. melanogaster* and *D. simulans*, whereas the majority of Alu elements present in the human genome are also found in the genome of the chimpanzee, despite a slightly longer divergence time between these two species compared with the two drosophilids. Although the vast majority of elements in the *Drosophila* species are polymorphic, unlike in the human system this information cannot be used to make inferences about the evolutionary history of a transposable element family.

There are several limitations to the methods employed to investigate the evolution of transposable elements. Obtaining complete sets of elements belonging to young Alu subfamilies is limited by the accuracy of the search methods employed. Although relatively relaxed search criteria are employed, more divergent elements are more likely to remain undetected, leading to an underestimation of the number of elements in a family and the total number of mutations. In other studies, where mutational data is used to estimate the age of the family, this may also lead to an underestimation of the time of origin of a particular subfamily. In addition, it is difficult to determine in which cases mutations are shared due to transposition rather than parallel mutation, which is further complicated by the potential for gene conversion. This is particularly true for CpG dinucleotides, at which transition mutations occur at six times the rate of mutations at other bases (Xing et al. 2004). In simulating the evolution of young Alu subfamilies, several assumptions are made. For example, it is assumed that all active elements within a subfamily produce daughter elements at the same rate, i.e. each active element is equally likely to provide the template for the next retrotransposition event. It is also assumed that the mutation rate, aside from the increase at CpG positions, is equal for each base in the element, and in all elements of the family. The founder gene is also constrained in the simulations, which does not allow for a master gene phylogeny to be produced with mutations accumulating in the master gene over time. In the investigation of horizontal transfer in *Drosophila*, several factors can reduce the certainty of the inference of horizontal transfer based on particular observations. For example, phylogenetic incongruence can provide a good indication of horizontal transfer, but it is possible to obtain incongruent phylogenies even when transfer has not occurred, thus potentially resulting in false positive results. In addition, as congruent phylogenies can also be generated in cases where horizontal transfer has occurred, depending on the relationship between the species involved and whether elements not descended from the transferred element are present in the recipient species, this may result in genuine cases of horizontal transfer being poorly supported or even undetected. A further limitation in the horizontal transfer investigation involves the inference of transfer events between extremely closely related species, such as *D. simulans* and *D. sechellia*, or *D. pseudoobscura* and *D. persimilis*, as both of these pairs of species diverged only around two million years ago. As a consequence, many of the transposable elements present in both species of one of these pairs are identical to each other by chance, simply as a result of the short divergence time in which mutations could have happened. Particularly in transposable element families and species which have high copy numbers, observations of identical elements are expected between such closely-related species. Therefore, this observation cannot be taken as evidence for horizontal transfer, and phylogenetic incongruence is almost always observed, with elements from the two species jumbled within a single clade. As a result, it is not possible to confidently infer horizontal transfer events between these species, and therefore the total number of horizontal transfer events may be underestimated. It is possible that horizontal transfer among these species may indeed be more common, as they are perhaps more likely to hybridise or to overlap in both geographical location and ecological niche. Geographical overlap between the putative donor and recipient species is assumed to be essential for horizontal transfer to occur. The results showed that in all but one case where horizontal transfer was inferred, there was indeed geographical overlap between the species involved. In the final case, a putative transfer of Minos involving *D. yakuba* and *D. mojavensis*, it is possible that an intermediate species was involved. Therefore, the assumption that geographical overlap is required may be invalid, as it may be possible for intermediate vectors to carry the transposable element from one species to the other. For example, the donor species may overlap with an intermediate species in part of its range, and that intermediate species may overlap with the recipient in a different part of its range. This appears to have occurred in the case of Minos. Furthermore, horizontal transfer may have occurred from a donor species for which the complete sequenced genome is unavailable, which overlaps with the recipient species, but which is closely-related to a species for which the complete genome sequence is available, but which does not overlap geographically with the recipient. As a result of lack of geographical overlap between the two species for which genome sequence data is available, horizontal transfer may be overlooked in this case where it has actually occurred. Once again this is indicative of the strict requirements that must be met in order for horizontal transfer to be confidently inferred, which may result in an underestimate of the actual number of cases of horizontal transfer of transposable elements. This does not, however, appear to be a serious limitation, as in all but one case, geographical overlap was observed between putative donor and recipient species. This may be a result of closely-related species tending to occupy similar geographical ranges.

The limitations of comparing the smallest divergence between transposable elements of the same family in different species with host genes in those species is exemplified by several of the families examined which are found in closely-related species, such as Paris, S2 and Diver2. For example, the DNA transposon family Paris is implicated to have been involved in horizontal transfer through comparison of elements in *D. persimilis* and *D. mojavensis*, where the smallest divergence between elements is less than that for the *Adh* coding region in these species, at 0.180, compared with 0.207 between the *Adh* genes. However, given the divergence observed is consistent with a relatively ancient horizontal transfer event, any such event is likely to have involved the ancestor of *D. pseudoobscura* and *D. persimilis*, which diverged around two million years ago. However, the divergence between Paris elements in *D. pseudoobscura* and *D. mojavensis* is greater than for *Adh*, although it is smaller than the divergence between *Adh* in *D. persimilis* and *D. mojavensis*. Under the stringent criteria of requiring the divergence between elements to be less than that between host genes, this potential case of horizontal transfer would not be detected if, for example, the genome sequence for *D. pseudoobscura* was available, but the sequence for *D. persimilis* was not. This case illustrates the need to assess the evidence available for each individual case, rather than broadly applying criteria to determine a cut-off for when horizontal transfer is a likely explanation for the observation, or is discounted. Using the criterion of observing divergence between elements smaller than that between host genes, particularly as host genes can perhaps be assumed to be under greater constraint than transposable element sequences, may lead to further underestimation of the number of cases of horizontal transfer. The limitations of the methods used to infer horizontal transfer do have a tendency to reduce the resolution of the analyses to detect horizontal transfer events, rather than increase the likelihood of false positive results. Therefore, it can be assumed that horizontal transfer is indeed a frequent occurrence for transposable element families in *Drosophila*, and may be even more common than the results presented here suggest.

To further the investigations described here, several lines of enquiry could be followed. Firstly, the investigation of the evolution of Alu elements in humans could be extended to include a greater number of young subfamilies, or to include members of the older AluS and AluJ subfamilies. Such an investigation would require different methods to be employed, as it is much more difficult to confidently assign older Alu elements to particular subfamilies, or to elucidate mutations which have been propagated by the activity of secondary source elements rather than parallel mutation, which itself is a more frequent observation when considering older families. Furthermore, insertion polymorphism would not be available as a source of information, as most of the older Alu subfamilies are currently inactive, with some exceptions, which are only active at a low level (Johanning et al. 2003). Inclusion of a greater number of young Alu subfamilies might allow broader patterns or common trends to be observed in the amplification dynamics of Alu elements of this type. For example, it might be possible to infer the estimated average proportion of young Alu subfamilies which have been contributed by the activity of secondary source elements. As improved genome assemblies for both the gorilla and orangutan become available, these sequences could be used to extend the investigation of young Alu subfamilies. For example, it would be possible to determine whether the DB3 locus, purported to be the master gene of the AluYh3a3 subfamily, is present and active in the gorilla genome, which could add further support to the hypothesis of gene conversion leading to inactivation of this master gene in humans.

The program used to simulate the evolution of young Alu subfamilies could be modified to allow for different active elements in the family to have different levels of activity. As suggested by another model of young Alu subfamily amplification (Cordaux et al. 2004), it may be that there is a primary master gene which contributes the majority of elements in a subfamily, with several other secondary elements contributing relatively few. This hypothesis does indeed fit with the data observed in at least some cases, such as for AluYi6, where although both the results of the simulations and analysis of polymorphism and mutation data suggest the activity of many source elements, the element corresponding to the consensus sequence appears to have produced the majority of elements in humans. In chimpanzees, an element with a different sequence, possessing two mutations from the consensus sequences, appears to have produced around a third of the total elements in the subfamily. A more accurate estimation of the range of values for the rate of retrotransposition (pT) and the activation rate (pA) could be obtained by testing values in smaller increments rather than on a logarithmic scale.

As the genomes of more species of the *Drosophila* genus, and other related species, become available, the investigation of horizontal transfer could be extended to include these species. As 52% of transposable element families investigated were found to have undergone at least one horizontal transfer event within the twelve species for which the sequenced genome is currently available, it is likely that many further horizontal transfer events would be identified were the investigation extended to include other species. Furthermore, cases are likely to be identified for the families for which there is no evidence for horizontal transfer among the current twelve genomes. This may perhaps lead to the suggestion that horizontal transfer has occurred at least once for the vast majority, if not all, transposable element families in *Drosophila*. The availability of additional genome sequences is also likely to elucidate directions of transfers which cannot currently be inferred, for example, where a transposable element family is only observed in the genomes of the two species between which the transfer is believed to have occurred. The species involved in horizontal transfer events may also be more accurately determined, as it is possible that a species very closely related to one of the twelve species for which the complete genome sequence is available may have acted as a donor species, whereas with the current data the species with the sequenced genome would be assigned as the donor. Investigation of horizontal transfer using further species would therefore allow for a more accurate assessment of the relative frequency of horizontal transfer involving particular species or pairs of species, and may allow realistic estimates of the age of horizontal transfer events to be made in some cases. For example, if a recipient species has a closely-related sister species which diverged one million years ago, and the transposable element in question was absent from this species, the transfer could be dated to over a million years ago. In addition, further relationships might allow for an upper limit on the age of the event to be assigned. It is also possible that in cases where the evidence supporting horizontal transfer is relatively weak, this may be attributed to the donor species not being represented in the twelve genomes. Further genomes may therefore provide additional support to many of the current inferences of horizontal transfer presented here.

References

Abrusan, G. & Krambeck, H.J. 2006a. Competition may determine the diversity of transposable elements. *Theoretical Population Biology*, 70, (3) 364-375

Abrusan, G. & Krambeck, H.J. 2006b. The distribution of L1 and Alu retroelements in relation to GC content on human sex chromosomes is consistent with the ectopic recombination model. *Journal of Molecular Evolution*, 63, (4) 484-492

Ackerman, H., Udalova, I., Hull, J., & Kwiatkowski, D. 2002. Evolution of a polymorphic regulatory element in interferon-gamma through transposition and mutation. *Molecular Biology and Evolution*, 19, (6) 884-890

Aleman, C., Roy-Engel, A.M., Shaikh, T.H., & Deininger, P.L. 2000. Cis-acting influences on Alu RNA levels. *Nucleic Acids Research*, 28, (23) 4755-4761

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., & Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25, (17) 3389-3402

An, H.J., Lee, D., Lee, K.H., & Bhak, J. 2004. The association of Alu repeats with the generation of potential AU-rich elements (ARE) at 3 ' untranslated regions. *Bmc Genomics*, 5, (1) 97

Aquadro, C.F., Lado, K.M., & Noon, W.A. 1988. The rosy region of *Drosophila melanogaster* and *Drosophila simulans*. I. Contrasting levels of naturally occurring DNA restriction map variation and divergence. *Genetics*, 119, (4) 875-888

Aravin, A.A., Naumova, N.M., Tulin, A.V., Vagin, V.V., Rozovsky, Y.M., & Gvozdev, V.A. 2001. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Current Biology*, 11, (13) 1017-1027

Arca, B. & Savakis, C. 2000. Distribution of the transposable element Minos in the genus *Drosophila*. *Genetica*, 108, (3) 263-267

Arkhipova, I.R. 1995. Complex patterns of transcription of a *Drosophila* retrotransposon in vivo and in vitro by RNA polymerases II and III. *Nucleic Acids Research*, 23, (21) 4480-4487

Ashburner, M. & Novitski, E. 1976. *The Genetics and Biology of Drosophila* Academic Press, London.

Austin, S. & Styles, P. 2009a. Nobel_I: A Bel-like LTR retrotransposon in *Drosophila persimilis*. *Repbase Reports*, 9, (6) 1151

Austin, S. & Styles, P. 2009b. Nobel_LTR: LTR sequence flanking Nobel_I in *Drosophila persimilis*. *Repbase Reports*, 9, (6) 1152

Bachellier, S., Clement, J.M., & Hofnung, M. 1999. Short palindromic repetitive DNA elements in enterobacteria: a survey. *Research in Microbiology*, 150, (9-10) 627-639

Bannert, N. & Kurth, R. 2004. Retroelements and the human genome: new perspectives on an old relation. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl 2, 14572-14579

Bartolome, C., Bello, X., & Maside, X. 2009. Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes. *Genome Biology*, 10, (2) R22

Batzer, M.A. & Deininger, P.L. 2002. Alu repeats and human genomic diversity. *Nature Reviews Genetics*, 3, (5) 370-379

Batzer, M.A., Deininger, P.L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C.M., Schmid, C.W., Zietkiewicz, E., & Zuckerkandl, E. 1996. Standardized nomenclature for Alu repeats. *Journal of Molecular Evolution*, 42, (1) 3-6

Belle, E.M.S., Webster, M.T., & Eyre-Walker, A. 2005. Why are young and old repetitive elements distributed differently in the human genome? *Journal of Molecular Evolution*, 60, (3) 290-296

Bergman, C.M. & Bensasson, D. 2007. Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, 104, (27) 11340-11345

Bergman, C.M., Quesneville, H., Anxolabehere, D., & Ashburner, M. 2006. Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biology*, 7, (11) R112

Bingham, P.M., Kidwell, M.G., & Rubin, G.M. 1982. The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. *Cell*, 29, (3) 995-1004

Boeke, J.D. 1997. LINEs and Alus - The polyA connection. *Nature Genetics*, 16, (1) 6-7

Bogerd, H.P., Wiegand, H.L., Hulme, A.E., Garcia-Perez, J.L., O'Shea, K.S., Moran, J.V., & Cullen, B.R. 2006. Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. *Proceedings of the National Academy of Sciences of the United States of America*, 103, (23) 8780-8785

Brezinsky, L., Wang, G.V., Humphreys, T., & Hunt, J. 1990. The transposable element Uhu from Hawaiian *Drosophila*--member of the widely dispersed class of Tc1-like transposons. *Nucleic Acids Research*, 18, (8) 2053-2059

Brookfield, J.F., Montgomery, E., & Langley, C.H. 1984. Apparent absence of transposable elements related to the P elements of D. melanogaster in other species of Drosophila. *Nature*, 310, (5975) 330-332

Brookfield, J.F.Y. 2001. Selection on Alu sequences? *Current Biology*, 11, (22) R900-R901

Brookfield, J.F.Y. & Johnson, L.J. 2006. The evolution of mobile DNAs: When will transposons create phylogenies that look as if there is a master gene? *Genetics*, 173, (2) 1115-1123

Brunet, F., Godin, F., Bazin, C., & Capy, P. 1999. Phylogenetic analysis of Mos1-like transposable elements in the Drosophilidae. *Journal of Molecular Evolution*, 49, (6) 760-768

Bull, T.J., Sidi-Boumedine, K., Mcminn, E.J., Stevenson, K., Pickup, R., & Hermon-Taylor, J. 2003. Mycobacterial interspersed repetitive units (MIRU) differentiate *Mycobacterium avium* subspecies *paratuberculosis* from other species of the *Mycobacterium avium* complex. *Molecular and Cellular Probes*, 17, (4) 157-164

Carroll, M.L., Owens, S., Roy-Engel, A.M., Nguyen, S., & Batzer, M.A. 2002. Human genomic diversity and an analysis of the Alu Ya5 and Yb8 subfamilies. *Faseb Journal*, 16, (4) A542

Charlesworth, B., Jarne, P., & Assimacopoulos, S. 1994. The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. III. Element abundances in heterochromatin. *Genetical Research*, 64, (3) 183-197

Charlesworth, B. & Langley, C.H. 1989. The population genetics of *Drosophila* transposable elements. *Annual Review of Genetics*, 23, 251-287

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., & Thompson, J.D. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, 31, (13) 3497-3500

Chiu, Y.L., Witkowska, H.E., Hall, S.C., Santiago, M., Soros, V.B., Esnault, C., Heidmann, T., & Greene, W.C. 2006. High-molecular-mass APOBEC3G complexes restrict Alu retrotransposition. *Proceedings of the National Academy of Sciences of the United States of America*, 103, (42) 15588-15593

Chung, H., Bogwitz, M.R., McCart, C., Andrianopoulos, A., Ffrench-Constant, R.H., Batterham, P., & Daborn, P.J. 2007. Cis-regulatory elements in the Accord retrotransposon result in tissue-specific expression of the *Drosophila melanogaster* insecticide resistance gene *cyp6g1*. *Genetics*, 175, (3) 1071-1077

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., Pollard, D.A., Sackton, T.B., Larracuente, A.M., Singh, N.D., Abad, J.P., Abt, D.N., Adryan, B., Aguade, M., Akashi, H., Anderson, W.W., Aquadro, C.F., Ardell, D.H., Arguello, R., Artieri, C.G., Barbash, D.A., Barker, D., Barsanti, P., Batterham, P., Batzoglou, S., Begun, D., Bhutkar, A., Blanco, E., Bosak, S.A., Bradley, R.K., Brand, A.D., Brent, M.R., Brooks, A.N., Brown, R.H., Butlin, R.K., Caggese, C., Calvi, B.R., Bernardo de, C.A., Caspi, A., Castrezana, S., Celniker, S.E., Chang, J.L., Chapple, C., Chatterji, S., Chinwalla, A., Civetta, A., Clifton, S.W., Comeron, J.M., Costello, J.C., Coyne, J.A., Daub, J., David, R.G., Delcher, A.L., Delehaunty, K., Do, C.B., Ebling, H., Edwards, K., Eickbush, T., Evans, J.D., Filipski, A., Findeiss, S., Freyhult, E., Fulton, L., Fulton, R., Garcia, A.C., Gardiner, A., Garfield, D.A., Garvin, B.E., Gibson, G., Gilbert, D., Gnerre, S., Godfrey, J., Good, R., Gotea, V., Gravely, B., Greenberg, A.J., Griffiths-Jones, S., Gross, S., Guigo, R., Gustafson, E.A., Haerty, W., Hahn, M.W., Halligan, D.L., Halpern, A.L., Halter, G.M., Han, M.V., Heger, A., Hillier, L., Hinrichs, A.S., Holmes, I., Hoskins, R.A., Hubisz, M.J., Hultmark, D., Huntley, M.A., Jaffe, D.B., Jagadeeshan, S., Jeck, W.R., Johnson, J., Jones, C.D., Jordan, W.C., Karpen, G.H., Kataoka, E., Keightley, P.D., Kheradpour, P., Kirkness, E.F., Koerich, L.B., Kristiansen, K., Kudrna, D., Kulathinal, R.J., Kumar, S., Kwok, R., Lander, E., Langley, C.H., Lapoint, R., Lazzaro, B.P., Lee, S.J., Levesque, L., Li, R., Lin, C.F., Lin, M.F., Lindblad-Toh, K., Llopart, A., Long, M., Low, L., Lozovsky, E., Lu, J., Luo, M., Machado, C.A., Makalowski, W., Marzo, M., Matsuda, M., Matzkin, L., McAllister, B., McBride, C.S., McKernan, B., McKernan, K., Mendez-Lago, M., Minx, P., Mollenhauer, M.U., Montooth, K., Mount, S.M., Mu, X., Myers, E., Negre, B., Newfeld, S., Nielsen, R., Noor, M.A., O'Grady, P., Pachter, L., Papaceit, M., Parisi, M.J., Parisi, M., Parts, L., Pedersen, J.S., Pesole, G., Phillippy, A.M., Ponting, C.P., Pop, M., Porcelli, D., Powell, J.R., Prohaska, S., Pruitt, K., Puig, M., Quesneville, H., Ram, K.R., Rand, D., Rasmussen, M.D., Reed, L.K., Reenan, R., Reily, A., Remington, K.A., Rieger, T.T., Ritchie, M.G., Robin, C., Rogers, Y.H., Rohde, C., Rozas, J., Rubenfield, M.J., Ruiz, A., Russo, S., Salzberg, S.L., Sanchez-Gracia, A., Saranga, D.J., Sato, H., Schaeffer, S.W., Schatz, M.C., Schlenke, T., Schwartz, R., Segarra, C., Singh, R.S., Sirot, L., Sirota, M., Sisneros, N.B., Smith, C.D., Smith, T.F., Spieth, J., Stage, D.E., Stark, A., Stephan, W., Strausberg, R.L., Strempel, S., Sturgill, D., Sutton, G., Sutton, G.G., Tao, W., Teichmann, S., Tobari, Y.N., Tomimura, Y., Tsolas, J.M., Valente, V.L., Venter, E., Venter, J.C., Vicario, S., Vieira, F.G., Vilella, A.J., Villasante, A., Walenz, B., Wang, J., Wasserman, M., Watts, T., Wilson, D., Wilson, R.K., Wing, R.A., Wolfner, M.F., Wong, A., Wong, G.K., Wu, C.I., Wu, G., Yamamoto, D., Yang, H.P., Yang, S.P., Yorke, J.A., Yoshida, K., Zdobnov, E., Zhang, P., Zhang, Y., Zimin, A.V., Baldwin, J., Abdouelleil, A., Abdulkadir, J., Abebe, A., Abera, B., Abreu, J., Acer, S.C., Aftuck, L., Alexander, A., An, P., Anderson, E., Anderson, S., Arachi, H., & Azer, M. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450, (7167) 203-218

Clark, J.B. & Kidwell, M.G. 1997. A phylogenetic perspective on P transposable element evolution in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, 94, (21) 11428-11433

Clark, J.B., Kim, P.C., & Kidwell, M.G. 1998. Molecular evolution of P transposable elements in the genus *Drosophila*. III. The *melanogaster* species group. *Molecular Biology and Evolution*, 15, (6) 746-755

Comas, D., Plaza, S., Calafell, F., Sajantila, A., & Bertranpetit, J. 2001. Recent insertion of an Alu element within a polymorphic human-specific Alu insertion. *Molecular Biology and Evolution*, 18, (1) 85-88

Cordaux, R., Hedges, D.J., & Batzer, M.A. 2004. Retrotransposition of Alu elements: how many sources? *Trends in Genetics*, 20, (10) 464-467

Cordaux, R., Hedges, D.J., Herke, S.W., & Batzer, M.A. 2006a. Estimating the retrotransposition rate of human Alu elements. *Gene*, 373, 134-137

Cordaux, R., Lee, J., Dinoso, L., & Batzer, M.A. 2006b. Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. *Gene*, 373, 138-144

Crosby, M.A., Goodman, J.L., Strelets, V.B., Zhang, P., & Gelbart, W.M. 2007. FlyBase: genomes by the dozen. *Nucleic Acids Research*, 35, (Database issue) D486-D491

Daniels, S.B., Peterson, K.R., Strausbaugh, L.D., Kidwell, M.G., & Chovnick, A. 1990. Evidence for horizontal transmission of the P transposable element between *Drosophila* species. *Genetics*, 124, (2) 339-355

de Almeida, L.M. & Carareto, C.M. 2005. Multiple events of horizontal transfer of the Minos transposable element between *Drosophila* species. *Molecular Phylogenetics and Evolution*, 35, (3) 583-594

de la Chaux, N. & Wagner, A. 2009. Evolutionary dynamics of the LTR retrotransposons roo and rooA inferred from twelve complete *Drosophila* genomes. *Bmc Evolutionary Biology*, 9, 205

de Setta, N., Loreto, E.L., & Carareto, C.M. 2007. Is the evolutionary history of the O-type P element in the *saltans* and *willistoni* groups of *Drosophila* similar to that of the canonical P element? *Journal of Molecular Evolution*, 65, (6) 715-724

Deininger, P.L., Batzer, M.A., Hutchison, C.A., & Edgell, M.H. 1992. Master genes in mammalian repetitive DNA amplification. *Trends in Genetics*, 8, (9) 307-311

Depra, M., Valente, V.L., Margis, R., & Loreto, E.L. 2009. The hobo transposon and hobo-related elements are expressed as developmental genes in *Drosophila*. *Gene*, 448, (1) 57-63

Dewannieux, M., Esnault, C., & Heidmann, T. 2003. LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics*, 35, (1) 41-48

Dewannieux, M. & Heidmann, T. 2005. Role of poly(A) tail length in Alu retrotransposition. *Genomics*, 86, (3) 378-381

Eickbush, T.H. & Furano, A.V. 2002. Fruit flies and humans respond differently to retrotransposons. *Current Opinion in Genetics & Development*, 12, (6) 669-674

Fablet, M., Lerat, E., Rebollo, R., Horard, B., Burlet, N., Martinez, S., Brasset, E., Gilson, E., Vaury, C., & Vieira, C. 2009. Genomic environment influences the dynamics of the tirant LTR retrotransposon in *Drosophila*. *Faseb Journal*, 23, (5) 1482-1489

Fablet, M., Rebollo, R., Biemont, C., & Vieira, C. 2007. The evolution of retrotransposon regulatory regions and its consequences on the *Drosophila melanogaster* and *Homo sapiens* host genomes. *Gene*, 390, (1-2) 84-91

Franz, G. & Savakis, C. 1991. Minos, a new transposable element from *Drosophila hydei,* is a member of the Tc1-like family of transposons. *Nucleic Acids Research*, 19, (23) 6646

Fukuuchi, A., Nagamura, Y., Yaguchi, H., Ohkura, N., Obara, T., & Tsukada, T. 2006. A whole MEN1 gene deletion flanked by Alu repeats in a family with multiple endocrine neoplasia type 1. *Japanese Journal of Clinical Oncology*, 36, (11) 739-744

Garcia Guerreiro, M.P., Chavez-Sandoval, B.E., Balanya, J., Serra, L., & Fontdevila, A. 2008. Distribution of the transposable elements bilbo and gypsy in original and colonizing populations of *Drosophila subobscura*. *Bmc Evolutionary Biology*, 8, 234

Garcia-Planells, J., Paricio, N., Clark, J.B., de, F.R., & Kidwell, M.G. 1998. Molecular evolution of P transposable elements in the genus *Drosophila*. II. The *obscura* species group. *Journal of Molecular Evolution*, 47, (3) 282-291

Gasior, S.L., Preston, G., Hedges, D.J., Gilbert, N., Moran, J.V., & Deininger, P.L. 2007. Characterization of pre-insertion loci of de novo L1 insertions. *Gene*, 390, (1-2) 190-198

Gibbons, R., Dugaiczyk, L.J., Girke, T., Duistermars, B., Zielinski, R., & Dugaiczky, A. 2004. Distinguishing humans from great apes with AluYb8 repeats. *Journal of Molecular Biology*, 339, (4) 721-729

Gilbert, C., Pace, J.K., & Feschotte, C. 2009. Horizontal SPINning of transposons. *Communicative and Integrative Biology*, 2, (2) 117-119

Glazko, G.V. & Nei, M. 2003. Estimation of divergence times for major lineages of primate species. *Molecular Biology and Evolution*, 20, (3) 424-434

Gombart, A.F., Saito, T., & Koeffler, H.P. 2009. Exaptation of an ancient Alu short interspersed element provides a highly conserved vitamin D-mediated innate immune response in humans and primates. *Bmc Genomics*, 10, 321

Gonzalez, J., Macpherson, J.M., & Petrov, D.A. 2009. A recent adaptive transposable element insertion near highly conserved developmental loci in *Drosophila melanogaster*. *Molecular Biology and Evolution*, 26, (9) 1949-1961

Granzotto, A., Lopes, F.R., Lerat, E., Vieira, C., & Carareto, C.M. 2009. The evolutionary dynamics of the Helena retrotransposon revealed by sequenced *Drosophila* genomes. *Bmc Evolutionary Biology*, 9, 174

Greally, J.M. 2002. Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 99, (1) 327-332

Guindon, S., Lethiec, F., Duroux, P., & Gascuel, O. 2005. PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Research*, 33, (Web Server issue) W557-W559

Hagemann, S., Haring, E., & Pinsker, W. 1996. Repeated horizontal transfer of P transposons between *Scaptomyza pallida* and *Drosophila bifasciata*. *Genetica*, 98, (1) 43-51

Han, K.D., Xing, J.C., Wang, H., Hedges, D.J., Garber, R.K., Cordaux, R., & Batzer, M.A. 2005. Under the genomic radar: The Stealth model of Alu amplification. *Genome Research*, 15, (5) 655-664

Haring, E., Hagemann, S., & Pinsker, W. 1998. Transcription and splicing patterns of M- and O-type P elements in *Drosophila bifasciata*, *D. helvetica*, and *Scaptomyza pallida*. *Journal of Molecular Evolution*, 46, (5) 542-551

Haring, E., Hagemann, S., & Pinsker, W. 2000. Ancient and recent horizontal invasions of Drosophilids by P elements. *Journal of Molecular Evolution*, 51, (6) 577-586

Hasler, J. & Strub, K. 2006. Alu elements as regulators of gene expression. *Nucleic Acids Research*, 34, (19) 5491-5497

Hedges, D.J., Callinan, P.A., Cordaux, R., Xing, J.C., Barnes, E., & Batzer, M.A. 2004. Differential Alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Research*, 14, (6) 1068-1075

Hedges, D.J., Cordaux, R., Xing, J.C., Witherspoon, D.J., Rogers, A.R., Jorde, L.B., & Batzer, M.A. 2005. Modeling the amplification dynamics of human Alu retrotransposons. *Plos Computational Biology*, 1, (4) 333-340

Heredia, F., Loreto, E.L., & Valente, V.L. 2004. Complex evolution of gypsy in Drosophilid species. *Molecular Biology and Evolution*, 21, (10) 1831-1842

Holland, L. Z. 10-4-2006. A SINE in the genome of the cephalochordate amphioxus is an Alu element. International Journal of Biological Sciences 2[2], 61-65.

Johanning, K., Stevenson, C.A., Oyeniran, O.O., Gozal, Y.M., Roy-Engel, A.M., Jurka, J., & Deininger, P.L. 2003. Potential for retroposition by old Alu subfamilies. *Journal of Molecular Evolution*, 56, (6) 658-664

Johnson, L.J. & Brookfield, J.F.Y. 2006. A test of the master gene hypothesis for interspersed repetitive DNA sequences. *Molecular Biology and Evolution*, 23, (2) 235-239

Jordan, I.K., Matyunina, L.V., & McDonald, J.F. 1999. Evidence for the recent horizontal transfer of long terminal repeat retrotransposon. *Proceedings of the National Academy of Sciences of the United States of America*, 96, (22) 12621-12625

Jordan, I.K., Rogozin, I.B., Glazko, G.V., & Koonin, E.V. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics*, 19, (2) 68-72

Jurka, J. 2004. Evolutionary impact of human Alu repetitive elements. *Current Opinion in Genetics & Development*, 14, (6) 603-608

Jurka, J. & Gentles, A.J. 2006. Origin and diversification of minisatellites derived from human Alu sequences. *Gene*, 365, 21-26

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, (1-4) 462-467

Jurka, J. & Klonowski, P. 1996. Integration of retroposable elements in mammals: Selection of target sites. *Journal of Molecular Evolution*, 43, (6) 685-689

Jurka, J., Kohany, O., Pavlicek, A., Kapitonov, V.V., & Jurka, M.V. 2004. Duplication, coclustering, and selection of human Alu retrotransposons. *Proceedings of the National Academy of Sciences of the United States of America*, 101, (5) 1268-1272

Jurka, J., Krnjajic, M., Kapitonov, V.V., Stenger, J.E., & Kokhanyy, O. 2002. Active Alu elements are passed primarily through paternal germlines. *Theoretical Population Biology*, 61, (4) 519-530

Jurka, J. & Milosavljevic, A. 1991. Reconstruction and analysis of human Alu genes. *Journal of Molecular Evolution*, 32, (2) 105-121

Kapitonov, V. & Jurka, J. 1996. The age of Alu subfamilies. *Journal of Molecular Evolution*, 42, (1) 59-65

Kapitonov, V.V. & Jurka, J. 2002a. GYPSY9, an old family of Gypsy-like endogenous retroviruses in *Drosophila*. *Repbase Reports*, 2, (11) 10

Kapitonov, V.V. & Jurka, J. 2002b. Looper1_DM, a family of Looper/PiggyBac DNA transposons in *D. melanogaster*. *Repbase Reports*, 2, (3) 6

Kapitonov, V.V. & Jurka, J. 2003a. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proceedings of the National Academy of Sciences of the United States of America*, 100, (11) 6569-6574

Kapitonov, V.V. & Jurka, J. 2003b. The esterase and PHD domains in CR1-like non-LTR retrotransposons. *Molecular Biology and Evolution*, 20, (1) 38-46

Kapitonov, V.V. & Jurka, J. 2007. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends in Genetics*, 23, (10) 521-529

Kass, D.H., Batzer, M.A., & Deininger, P.L. 1995. Gene conversion as a secondary mechanism of Short Interspersed Element (SINE) Evolution. *Molecular and Cellular Biology*, 15, (1) 19-25

Kass, D.H., Jamison, N., Mayberry, M.M., & Tecle, E. 2007. Identification of a unique Alu-based polymorphism and its use in human population studies. *Gene*, 390, (1-2) 146-152

Kazazian, H.H. 2004. Mobile elements: Drivers of genome evolution. *Science*, 303, (5664) 1626-1632

Kent, W.J. 2002. BLAT - The BLAST-like alignment tool. *Genome Research*, 12, (4) 656-664

Kidwell, M.G., Kidwell, J.F., & Sved, J.A. 1977. Hybrid dysgenesis in *Drosophila melanogaster*: a syndrome of aberrant traits including mutation, sterility and male recombination. *Genetics*, 86, (4) 813-833

Kidwell, M.G. & Lisch, D.R. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution*, 55, (1) 1-24

Kim, A., Terzian, C., Santamaria, P., Pelisson, A., Purd'homme, N., & Bucheton, A. 1994. Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, 91, (4) 1285-1289

Kimura, K. & Kidwell, M.G. 1994. Differences in P element population dynamics between the sibling species *Drosophila melanogaster* and *Drosophila simulans*. *Genetical Research*, 63, (1) 27-38

Kohany, O., Gentles, A.J., Hankus, L., & Jurka, J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics.*, 7, 474

Kotnova, A.P., Glukhov, I.A., Karpova, N.N., Salenko, V.B., Lyubomirskaya, N.V., & Ilyin, Y.V. 2007. Evidence for recent horizontal transfer of gypsy-homologous LTR-retrotransposon gtwin into *Drosophila erecta* followed by its amplification with multiple aberrations. *Gene*, 396, (1) 39-45

Kriegs, J. O., Churakov, G., Jurka, J., Brosius, J., and Schmitz, J. 1-4-2007. Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. Trends in Genetics 4, 158-161.

Kumar, S., Nei, M., Dudley, J., & Tamura, K. 2008. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics*, 9, (4) 299-306

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H.M., Yu, J., Wang, J., Huang, G.Y., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S.Z., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H.Q., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W.H., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J.R., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., & Morgan, M.J. 2001. Initial sequencing and analysis of the human genome. *Nature*, 409, (6822) 860-921

Lerat, E., Biemont, C., & Capy, P. 2000. Codon usage and the origin of P elements. *Molecular Biology and Evolution*, 17, (3) 467-468

Lerat, E., Rizzon, C., & Biemont, C. 2003. Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. *Genome Research*, 13, (8) 1889-1896

Li, T.H. & Schmid, C.W. 2001. Differential stress induction of individual Alu loci: implications for transcription and retrotransposition. *Gene*, 276, (1-2) 135-141

Li, T.H. & Schmid, C.W. 2004. Alu's dimeric consensus sequence destabilizes its transcripts. *Gene*, 324, 191-200

Llorens, J.V., Clark, J.B., Martinez-Garay, I., Soriano, S., de, F.R., & Martinez-Sebastian, M.J. 2008. Gypsy endogenous retrovirus maintains potential infectivity in several species of Drosophilids. *Bmc Evolutionary Biology*, 8, 302

Lobachev, K.S., Stenger, J.E., Kozyreva, O.G., Jurka, J., Gordenin, D.A., & Resnick, M.A. 2000. Inverted Alu repeats unstable in yeast are excluded from the human genome. *Embo Journal*, 19, (14) 3822-3830

Lohe, A.R., Moriyama, E.N., Lidholm, D.A., & Hartl, D.L. 1995. Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. *Molecular Biology and Evolution*, 12, (1) 62-72

Loreto, E.L., Carareto, C.M., & Capy, P. 2008. Revisiting horizontal transfer of transposable elements in *Drosophila*. *Heredity*, 100, (6) 545-554

Maity, T.S., Leonard, C.W., Rose, M.A., Fried, H.M., & Weeks, K.M. 2006. Compartmentalization directs assembly of the signal recognition particle. *Biochemistry*, 45, (50) 14955-14964

Makalowski, W. 2003. Not junk after all. *Science*, 300, (5623) 1246-1247

Malik, H.S., Burke, W.D., & Eickbush, T.H. 1999. The age and evolution of non-LTR retrotransposable elements. *Molecular Biology and Evolution*, 16, (6) 793-805

Martin-Campos, J.M., Comeron, J.M., Miyashita, N., & Aguade, M. 1992. Intraspecific and interspecific variation at the y-ac-sc region of *Drosophila simulans* and *Drosophila melanogaster*. *Genetics*, 130, (4) 805-816

Maruyama, K. & Hartl, D.L. 1991. Evolution of the transposable element mariner in *Drosophila* species. *Genetics*, 128, (2) 319-329

Maside, X., Bartolome, C., & Charlesworth, B. 2003. Inferences on the evolutionary history of the S-element family of *Drosophila melanogaster*. *Molecular Biology and Evolution*, 20, (8) 1183-1187

McClintock, B. 1953. Induction of Instability at Selected Loci in Maize. *Genetics*, 38, (6) 579-599

Medstrand, P., van de Lagemaat, L.N., & Mager, D.L. 2002. Retroelement distributions in the human genome: Variations associated with age and proximity to genes. *Genome Research*, 12, (10) 1483-1495

Mills, R.E., Bennett, E.A., Iskow, R.C., Luttig, C.T., Tsui, C., Pittard, W.S., & Devine, S.E. 2006. Recently mobilized Transposons in the human and chimpanzee Genomes. *American Journal of Human Genetics*, 78, (4) 671-679

Otieno, A.C., Carter, A.B., Hedges, D.J., Walker, J.A., Ray, D.A., Garber, R.K., Anders, B.A., Stoilova, N., Laborde, M.E., Fowlkes, J.D., Huang, C.H., Perodeau, B., & Batzer, M.A. 2004. Analysis of the human Alu Ya-lineage. *Journal of Molecular Biology*, 342, (1) 109-118

Pavlicek, A., Jabbari, K., Paces, J., Paces, V., Hejnar, J., & Bernardi, G. 2001. Similar integration but different stability of Alus and LINEs in the human genome. *Gene*, 276, (1-2) 39-45

Price, A.L., Eskin, E., & Pevzner, P.A. 2004. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Research*, 14, (11) 2245-2252

Quesneville, H., Bergman, C.M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., & Anxolabehere, D. 2005. Combined evidence annotation of transposable elements in genome sequences. *Plos Computational Biology*, 1, (2) 166-175

Ray, D.A. & Batzer, M.A. 2005. Tracking Alu evolution in New World primates. *Bmc Evolutionary Biology*, 5, 51

Retief, J.D. 2000. Phylogenetic analysis using PHYLIP. *Methods in Molecular Biology*, 132, 243-258

Rogers, J.H. 1985. The origin and evolution of retroposons. *International Review of Cytology-A Survey of Cell Biology*, 93, 187-279

Roy, A.M., Carroll, M.L., Kass, D.H., Nguyen, S.V., Salem, A.H., Batzer, M.A., & Deininger, P.L. 1999. Recently integrated human Alu repeats: finding needles in the haystack. *Genetica*, 107, (1-3) 149-161

Roy, A.M., Carroll, M.L., Nguyen, S.V., Salem, A.H., Oldridge, M., Wilkie, A.O.M., Batzer, M.A., & Deininger, P.L. 2000. Potential gene conversion and source genes for recently integrated Alu elements. *Genome Research*, 10, (10) 1485-1495

Roy-Engel, A.M., Carroll, M.L., El-Sawy, M., Salem, A.H., Garber, R.K., Nguyen, S.V., Deininger, P.L., & Batzer, M.A. 2002. Non-traditional Alu evolution and primate genomic diversity. *Journal of Molecular Biology*, 316, (5) 1033-1040

Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X., & Rozas, R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, 19, (18) 2496-2497

Salem, A.H., Kilroy, G.E., Watkins, W.S., Jorde, L.B., & Batzer, M.A. 2003a. Recently integrated Alu elements and human genomic diversity. *Molecular Biology and Evolution*, 20, (8) 1349-1361

Salem, A.H., Ray, D.A., Hedges, D.J., Jurka, J., & Batzer, M.A. 2005. Analysis of the human Alu Ye lineage. *Bmc Evolutionary Biology*, 5, (1) 18

Salem, A.H., Ray, D.A., Xing, J.C., Callinan, P.A., Myers, J.S., Hedges, D.J., Garber, R.K., Witherspoon, D.J., Jorde, L.B., & Batzer, M.A. 2003b. Alu elements and hominid phylogenetics. *Proceedings of the National Academy of Sciences of the United States of America*, 100, (22) 12787-12791

Sanchez-Gracia, A., Maside, X., & Charlesworth, B. 2005. High rate of horizontal transfer of transposable elements in *Drosophila*. *Trends in Genetics*, 21, (4) 200-203

Shen, M.R., Batzer, M.A., & Deininger, P.L. 1991. Evolution of the Master Alu Gene(s). *Journal of Molecular Evolution*, 33, (4) 311-320

Shimamura, M., Yasue, H., Ohshima, K., Abe, H., Kato, H., Kishiro, T., Goto, M., Munechika, I., & Okada, N. 1997. Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature*, 388, (6643) 666-670

Silva, J.C. & Kidwell, M.G. 2000. Horizontal transfer and selection in the evolution of P elements. *Molecular Biology and Evolution*, 17, (10) 1542-1557

Simmons, G.M. 1992. Horizontal transfer of hobo transposable elements within the *Drosophila melanogaster* species complex: evidence from DNA sequencing. *Molecular Biology and Evolution*, 9, (6) 1050-1060

Smalheiser, N.R. & Torvik, V.I. 2006. Alu elements within human mRNAs are probable microRNA targets. *Trends in Genetics*, 22, (10) 532-536

Styles, P. 2008a. Helitron1_Dmoj: a new family of helitrons in *Drosophila mojavensis*. *Repbase Reports*, 8, (9) 907

Styles, P. 2008b. Transib1_Dmoj: a new family of Transib elements in *D. mojavensis*. *Repbase Reports*, 8, (9) 908

Styles, P. 2008c. Transib1_Dwil: a new family of Transib elements in *Drosophila willistoni*. *Repbase Reports*, 8, (9) 909

Styles, P. 2009a. Gypsy_DG: A Gypsy-like family of LTR retrotransposons in *Drosophila grimshawi*. *Repbase Reports*, 9, (5) 962

Styles, P. 2009b. Gypsy_DG_LTR: The LTR sequence flanking Gypsy_DG_I in *Drosophila grimshawi*. *Repbase Reports*, 9, (5) 963

Styles, P. & Brookfield, J.F.Y. 2007. Analysis of the features and source gene composition of the AluYg6 subfamily of human retrotransposons. *Bmc Evolutionary Biology*, 7, 102

Styles, P. & Brookfield, J.F.Y. 2009. Source gene composition and gene conversion of the AluYh and AluYi lineages of retrotransposons. *Bmc Evolutionary Biology*, 9, 102

Syomin, B.V., Fedorova, L.I., Surkov, S.A., & Ilyin, Y.V. 2001. The endogenous Drosophila melanogaster retrovirus gypsy can propagate in *Drosophila hydei* cells. *Molecular and General Genetics*, 264, (5) 588-594

Syomin, B.V., Leonova, T.Y., & Ilyin, Y.V. 2002. Evidence for horizontal transfer of the LTR retrotransposon mdg3, which lacks an env gene. *Molecular Genetics and Genomics*, 267, (3) 418-423

Terzian, C., Ferraz, C., Demaille, J., & Bucheton, A. 2000. Evolution of the Gypsy endogenous retrovirus in the *Drosophila melanogaster* subgroup. *Molecular Biology and Evolution*, 17, (6) 908-914

van de Lagemaat, L.N., Gagnier, L., Medstrand, P., & Mager, D.L. 2005. Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. *Genome Research*, 15, (9) 1243-1249

Vassetzky, N.S., Ten, O.A., & Kramerov, D.A. 2003. B1 and related SINEs in mammalian genomes. *Gene*, 319, 149-160

Vidal, N.M., Ludwig, A., & Loreto, E.L. 2009. Evolution of Tom, 297, 17.6 and rover retrotransposons in Drosophilidae species. *Molecular Genetics and Genomics*, 282, (4) 351-362

Vieira, C. & Biemont, C. 2004. Transposable element dynamics in two sibling species: *Drosophila melanogaster* and *Drosophila simulans*. *Genetica*, 120, (1-3) 115-123

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown, S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins, F.S., Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis, E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S., Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T.A., Green, E.D., Gregory, S., Guigo, R., Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson, E.K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W.J., Kirby, A., Kolbe, D.L., Korf, I., Kucherlapati, R.S., Kulbokas, E.J., Kulp, D., Landers, T., Leger, J.P., Leonard, S., Letunic, I., LeVine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D.R., Mardis, E.R., Matthews, L., Mauceli, E., Mayer, J.H., McCarthy, M., McCombie, W.R., McLaren, S., McLay, K., McPherson, J.D., Meldrim, J., Meredith, B., Mesirov, J.P., Miller, W., Miner, T.L., Mongin, E., Montgomery, K.T., Morgan, M., Mott, R., Mullikin, J.C., Muzny, D.M., Nash, W.E., Nelson, J.O., Nhan, M.N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M.J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K.H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C.S., Poliakov, A., Ponce, T.C., Ponting, C.P., Potter, S., Quail, M., Reymond, A., Roe, B.A., Roskin, K.M., Rubin, E.M., Rust, A.G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M.S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J.B., Slater, G., Smit, A., Smith, D.R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J.P., Von Niederhausern, A.C., Wade, C.M., Wall, M., Weber, R.J., Weiss, R.B., Wendl, M.C., West, A.P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R.K., Winter, E., Worley, K.C., Wyman, D., Yang, S., Yang, S.P., Zdobnov, E.M., Zody, M.C., & Lander, E.S. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420, (6915) 520-562

Xing, J.C., Hedges, D.J., Han, K.D., Wang, H., Cordaux, R., & Batzer, M.A. 2004. Alu element mutation spectra: Molecular clocks and the effect of DNA methylation. *Journal of Molecular Biology*, 344, (3) 675-682

Zhi, D.G. 2007. Sequence correlation between neighboring Alu instances suggests post-retrotransposition sequence exchange due to Alu gene conversion. *Gene*, 390, (1-2) 117-121

Appendices

Appendix 1 - Source code for the Perl program ConsensusMatcher

```
#! usr/bin/perl

#Consensus matcher - matches any query sequence to any consensus sequence

print "\n";

print "CONSENSUS MATCHER\n";

print "\n";

$identity = 0;

print "Please enter the consensus sequence: \n";

$consensus = <STDIN>;
@consensus = split (", $consensus);

$copy_con = $consensus;
$copy_length = length $copy_con;
$consensus_length = $copy_length -1;

print "\nPlease enter the query sequence: \n";

$query = <STDIN>;

$copy_query = $query;
$query_length = length $copy_query;
$query_length = $query_length - 1;

# Explode into an array

@query = split (", $query);

if ($consensus eq $query) {
   print "\nMUTATIONS\n\nNo mutations found.\n";
   $identity = $copy_length;
} else {
  &mutation;
}

#Percent identity of query to consensus

$identity = $identity -1;

$percent_identity = ($identity/$consensus_length) * 100;

print "\n\n";
print "IDENTITY\n\n";
printf "Identity of query sequence to the consensus: %.2f%%",
$percent_identity;

print " ($identity\/$consensus_length).\n";

#GC content

$C_query = ($copy_query =~ tr/C//);
$G_query = ($copy_query =~ tr/G//);

$GC_query = $C_query + $G_query;
$GC_query = ($GC_query / $query_length) * 100;

$C_consensus = ($copy_con =~ tr/C//);
$G_consensus = ($copy_con =~ tr/G//);

$GC_consensus = $C_consensus + $G_consensus;
$GC_consensus = ($GC_consensus / $consensus_length) * 100;
```

```
print "\n\n";
print "GC CONTENT\n\n";

printf "GC content of the query sequence: %.2f%%.\n",
$GC_query;
printf "GC content of the consensus sequence: %.2f%%.\n",
$GC_consensus;

#Number of CpG dinucleotides

$CG_consensus = 0;

for ($position_con = 0; $position_con < length $copy_con; ++$position_con) {
   $base_con = substr ($copy_con, $position_con, 2);
       if ($base_con eq 'CG') {
                ++$CG_consensus;
        }
}

$CG_query = 0;

for ($position_query = 0; $position_query < length $copy_query; ++$position_query) {
   $base_query = substr ($copy_query, $position_query, 2);
       if ($base_query eq 'CG') {
                ++$CG_query;
        }
}

print "\n\n";
print "CpG DINUCLEOTIDES\n\n";

print "Number of CpG dinucleotides in the consensus sequence: $CG_consensus.\n";
print "Number of CpG dinucleotides in the query sequence: $CG_query.\n";

# SUBROUTINES

sub mutation {

print "\nMUTATIONS\n\n";

do {

$query_base = shift @query;
$con_base = shift @consensus;


if ($query_base eq $con_base) {
  $identity = $identity + 1;
} elsif (($query_base eq 'G') and ($con_base eq 'A')) {
      &AG;
} elsif (($query_base eq 'C') and ($con_base eq 'A')) {
      &AC;
} elsif (($query_base eq 'T') and ($con_base eq 'A')) {
      &AT;
} elsif (($query_base eq 'A') and ($con_base eq 'G')) {
      &GA;
} elsif (($query_base eq 'C') and ($con_base eq 'G')) {
      &GC;
} elsif (($query_base eq 'T') and ($con_base eq 'G')) {
      &GT;
} elsif (($query_base eq 'A') and ($con_base eq 'C')) {
      &CA;
} elsif (($query_base eq 'G') and ($con_base eq 'C')) {
      &CG;
} elsif (($query_base eq 'T') and ($con_base eq 'C')) {
      &CT;
} elsif (($query_base eq 'A') and ($con_base eq 'T')) {
      &TA;
} elsif (($query_base eq 'G') and ($con_base eq 'T')) {
      &TG;
} elsif (($query_base eq 'C') and ($con_base eq 'T')) {
      &TC;
} elsif ($query_base eq '-') {
      &deletion;
} elsif ($con_base eq '-') {
        &insertion;
}


} until ($query_base =~ /^\s*$/ );

}

sub AG {

    my $glue = "";
    my $consensus = join $glue, @consensus;
    my $sublength = length $consensus;
    my $position =  $copy_length - $sublength;

    print "A to G transition at position $position.\n";

    @consensus = split ('', $consensus);
}

sub AC {

    my $glue = "";
    my $consensus = join $glue, @consensus;
    my $sublength = length $consensus;
    my $position =  $copy_length - $sublength;

    print "A to C transversion at position $position.\n";

    @consensus = split ('', $consensus);
}

sub AT {

    my $glue = "";
    my $consensus = join $glue, @consensus;
    my $sublength = length $consensus;
    my $position =  $copy_length - $sublength;

    print "A to T transversion at position $position.\n";

    @consensus = split ('', $consensus);
}

sub GA {

    my $glue = "";
    my $consensus = join $glue, @consensus;
    my $sublength = length $consensus;
    my $position =  $copy_length - $sublength;

    print "G to A transition at position $position.\n";

    @consensus = split ('', $consensus);
}

sub GC {

    my $glue = "";
    my $consensus = join $glue, @consensus;
    my $sublength = length $consensus;
    my $position =  $copy_length - $sublength;

    print "G to C transversion at position $position.\n";

    @consensus = split ('', $consensus);
}

sub GT {

    my $glue = "";
    my $consensus = join $glue, @consensus;
    my $sublength = length $consensus;
    my $position =  $copy_length - $sublength;
    print "G to T transversion at position $position.\n";

    @consensus = split ('', $consensus);
}

sub CA {

    my $glue = "";
    my $consensus = join $glue, @consensus;
    my $sublength = length $consensus;
    my $position =  $copy_length - $sublength;

    print "C to A transversion at position $position.\n";

    @consensus = split ('', $consensus);
}

sub CG {

    my $glue = "";
    my $consensus = join $glue, @consensus;
    my $sublength = length $consensus;
    my $position =  $copy_length - $sublength;

    print "C to G transversion at position $position.\n";

    @consensus = split ('', $consensus);
}

sub CT {

    my $glue = "";
    my $consensus = join $glue, @consensus;
    my $sublength = length $consensus;
    my $position =  $copy_length - $sublength;

    print "C to T transition at position $position.\n";

    @consensus = split ('', $consensus);
}

sub TA {

    my $glue = "";
    my $consensus = join $glue, @consensus;
    my $sublength = length $consensus;
    my $position =  $copy_length - $sublength;

    print "T to A transversion at position $position.\n";

    @consensus = split ('', $consensus);
}

sub TG {

    my $glue = "";
    my $consensus = join $glue, @consensus;
    my $sublength = length $consensus;
    my $position =  $copy_length - $sublength;

    print "T to G transversion at position $position.\n";

    @consensus = split ('', $consensus);
}

sub TC {

    my $glue = "";
    my $consensus = join $glue, @consensus;
    my $sublength = length $consensus;
    my $position =  $copy_length - $sublength;

    print "T to C transition at position $position.\n";

    @consensus = split ('', $consensus);
```

```
}

sub deletion {

    my $glue = "";
    my $consensus = join $glue, @consensus;
    my $sublength = length $consensus;
    my $position =  $copy_length - $sublength;

    print "Deletion at position $position.\n";

    @consensus = split ('', $consensus);
}

sub insertion {

    my $glue = "";
    my $consensus = join $glue, @consensus;
    my $sublength = length $consensus;
    my $position =  $copy_length - $sublength;

    print "Insertion at position $position.\n";

    @consensus = split ('', $consensus);
}
```

Appendix 2 - Source code for the C++ program Subfamily Simulator, using example values for AluYh3a1 in chimpanzees

File 1: Subfamily.cpp

```
#include "Subfamily.h"
#include "Element.h"

#include <iostream>
#include <string>
#include <cmath>
#include <fstream>

using namespace std;

int a = 1;
int NTV = 0;
int NTS = 0;
int CTV = 0;
int CTS = 0;
int parallelMutations = 0;
int parallelSites = 0;
int variableSites = 0;
int sharedSourceMutations = 0;
double thetaResult;
double piResult;

double activeThreshold;
double retroThreshold;
int activeIncrement;
int retroIncrement;

bool parallelCTS[280];
bool parallelTS[280];
bool parallelTV1[280];
bool parallelTV2[280];
bool parallelCTV1[280];
bool parallelCTV2[280];
bool mutatedSites[280];

//Number of sources is initialised to 0
int sourceCount = 0;
int actives = 1;
long double totalTime = 0;
int ancestralMutations = 0;

Subfamily::Subfamily()
{
  Element element1;
  element1.designation = 1;
  element1.numberOfMutations = 0;
  element1.source =  false;
  element1.active = true;
  element1.parent = 0;
  element1.sequence =
"GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCAGAGGCGGGCGGATCATG
AGGTCAGGAGATCGAGACCATCCTGGCTAACACAGTGAAACCCCGCCTCTACTAAAAATACAAAAAAT
TAGCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGC
GTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACA
GAGCGAGACTCCGTCTC";
  elements.push_back(element1);

  //Number of elements created is initialised to 0
  elementsCreated = 0;
  sharedSourceMutations = 0;
  variableSites = 0;
  thetaResult = 0;
  piResult = 0;
  sequenceLength = element1.sequence.length();

  for (int i = 0; i < 280; i++)
  {
    parallelCTS[i] = 0;
    parallelTS[i] = 0;
    parallelCTV1[i] = 0;
    parallelCTV2[i] = 0;
    parallelTV1[i] = 0;
    parallelTV2[i] = 0;
    mutatedSites[i] = false;
  }

  for (elementsCreated = 0; elementsCreated < 72; )
  {
    eventDecision();
  }//end for

  for (int i=0; i<elements.size(); i++)
  {
    elements.at(i).printMe();
  }//end for

  for (int i=0; i<280; i++)
  {
    if (mutatedSites[i] == true)
    {
      variableSites++;
    }
  }
  thetaResult = theta();
  piResult = pi();
}//end constructor

//----------------------------------------------------------------

//RETROTRANSPOSITION FUNCTION

//----------------------------------------------------------------
void Subfamily::retrotransposition()
{
  if (elementsCreated == 0)
  {
    totalTime = 0;
  }
  int index;
  bool done = false;

  while (!done)
  {
    //get element at random index of vector
    index = rand() % elements.size();
    Element check = elements.at(index);
    if (check.active == true) done = true;
  }//end while

  Element& e1 = elements.at(index);

  a++;
  Element e2(&e1);                    //make a new element from e1
  if (e1.source == false) sourceCount++;    //if e1 is new source, increment sourceCount
  e1.source = true;
  sharedSourceMutations = sharedSourceMutations + e1.numberOfMutations;

  elements.push_back(e2);              //add new element to vector
  elementsCreated++;                  //increment elements created
}

//----------------------------------------------------------------

//COMPLETE FORWARD GENE CONVERSION FUNCTION

//----------------------------------------------------------------
void Subfamily::gcCompleteForward()
{
  //Generate random number between 0 and size of elements vector
  int f = rand() % elements.size();

  //Get the element at f
  Element& e1 = elements.at(f);
  //Make a new element from e1 and add it to the vector
  a++;
  Element e2 (&e1);

  elements.push_back(e2);
  elementsCreated++;
}
//----------------------------------------------------------------

//COMPLETE BACKWARD GENE CONVERSION FUNCTION

//----------------------------------------------------------------

void Subfamily::gcCompleteBackward(int index)
{
  vector<Element> v2;

  for (int i=0; i<elements.size(); i++)
  {
    if (index == i)
    {
      Element d = elements.at(i);
      continue;
    }
    Element e = elements.at(i);
    v2.push_back(e);

  }//end for

  elements = v2;

}

//----------------------------------------------------------------

// GENE CONVERSION DECISION FUNCTION

//----------------------------------------------------------------

void Subfamily::geneConversion()
{
  int conversion = rand() % 10;
  int index = rand() % elements.size();

  if (conversion <=2)
    gcCompleteForward();
  else if (conversion == 4)
    gcCompleteBackward(index);
  else if (conversion <=6)
    gcPartialAluSg();
  else
    gcPartialAluY();
}

//----------------------------------------------------------------
//Designation function
int Designation()
{
  a++;
  return a;
}

//----------------------------------------------------------------
//Activation function
bool Activate()
{
  double i;

  int j = rand() % 1000;
  i = (double) (j)/1000;

  if (i < activeThreshold)
  {
    actives++;
```

```
      return true;

   }
   else
      return false;
}
//----------------------------------------------------------------

// Partial Gene Conversion Functions

//----------------------------------------------------------------
void Subfamily::gcPartialAluSg()
{
   string AluSg =
"GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACG
AGGTCAGGAGTTCGAGACCAGCCTGGCCAACATGGTGAAACCCCGTCTCTACTAAAAATACAAAAATT
AGCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCGCTT
GAACCCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGA
GCGAGACTCCGTCTC";

   int tractStart = rand() % AluSg.length();

   int tractLength = (rand() % 50) + 50;

   int tractEnd = tractStart + tractLength;

   if (tractEnd > 280)
   {
      tractEnd = 280;
      tractLength = 280 - tractStart;
   }

   // cout << "The tract length is: " << tractLength << endl;

   //Generate random number between 0 and size of elements vector
   int f = rand() % elements.size();

   //Get the element at f
   Element& e1 = elements.at(f);

   int position;

   for (position = tractStart; position<=tractEnd;  position++)
   {
      e1.sequence[position] = AluSg[position];

   }
   if (tractStart <= 52 && tractEnd >= 52)
      ancestralMutations++;

   if (tractStart<= 142 && tractEnd >= 142)
      ancestralMutations++;

   if (tractStart <= 151 && tractEnd >= 151)
      ancestralMutations++;

   if (tractStart <= 172 && tractEnd >= 172)
      ancestralMutations++;

   if (tractStart<= 228 && tractEnd >= 228)
      ancestralMutations++;

   if (tractStart <= 270 && tractEnd >= 270)
      ancestralMutations++;
}

//----------------------------------------------------------------
void Subfamily::gcPartialAluY()
{
   string AluY =
"GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACG
AGGTCAGGAGATCGAGACCATCCTGGCTAACACGGTGAAACCCCGTCTCTACTAAAAATACAAAAATT
AGCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCG
TGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAG
AGCGAGACTCCGTCTC";
```

```
  int tractStart = rand() % AluY.length();

  int tractLength = (rand() % 50) + 50;

  int tractEnd = tractStart + tractLength;

  if (tractEnd > 280)
  {
    tractEnd = 280;
    tractLength = 280 - tractStart;
  }


//Generate random number between 0 and size of elements vector
  int f = rand() % elements.size();

//Get the element at f
  Element& e1 = elements.at(f);

  int position;

  for (position = tractStart; position<=tractEnd;  position++)
  {
    e1.sequence[position] = AluY[position];

  }
  if (tractStart <= 52 && tractEnd >= 52)
    ancestralMutations++;

  if (tractStart<= 142 && tractEnd >= 142)
    ancestralMutations++;

  if (tractStart <= 151 && tractEnd >= 151)
    ancestralMutations++;

  if (tractStart <= 172 && tractEnd >= 172)
    ancestralMutations++;

  if (tractStart<= 228 && tractEnd >= 228)
    ancestralMutations++;

  if (tractStart <= 270 && tractEnd >= 270)
    ancestralMutations++;
}

//----------------------------------------------------------------
//Choose element to mutate
void Subfamily::mutation()
{
  int index;
  bool done = false;

  while (!done)
  {
    //get element at random index of vector
    index = rand() % elements.size();
    //cout << "Random element: " << index << "\n";
    Element check = elements.at(index);
    if (check.designation != 1) done = true;
  }//end while

  Element& e1 = elements.at(index);

  e1.numberOfMutations++;
  e1.Mutation();
}
//----------------------------------------------------------------
double Subfamily::exponential()
{
  bool done = false;
  double s;

  while (!done)
  {
    int x = rand() % 1000;
    s = (double) (x)/1000;

    if (s != 0) done = true;

  }//end while

  double y = -log( s );

  return y;
 /}
}
//-----------------------------------------------------------------
void Subfamily::eventDecision()
{
  double y = exponential();
  double pM;

  if (elements.size() == 1)
  {
    pM = 0.00;
  }
  else
  {
    // pM = mutationThreshold;
    pM = 0.000000795;
  }

  pM = pM * (elementsCreated+1);

  double pT = retroThreshold;

  double averageT;
  double interceptT;
  double gradientT = 0.0000156;
  double variableT;
  double myrTime;

  averageT = pT;
  interceptT = averageT - 0.000001;

  myrTime = totalTime / 1000000;

  variableT = (gradientT * myrTime) + interceptT;

  pT = variableT;

  pT = pT * actives;

  double pG;

  if (actives <= 2)
  {
    pG = 0.00;
  }
  else
  {
   pG = 0.00000000;
  }

  pG = pG * (elementsCreated+1);

  double denominator = pT+pM+pG;
  double z = 1/denominator;

  double timeToEvent = y * z;

   totalTime = totalTime + timeToEvent;

  int x = rand() % 1000;
  double s = (double) (x)/1000;

  double pretrotransposition;
  pretrotransposition = pT;
  pretrotransposition = pretrotransposition/denominator;
```

```
   double pmutation;
   pmutation = pM;
   pmutation = pmutation/denominator;

   double pgeneconversion;
   pgeneconversion = pG;
   pgeneconversion = pgeneconversion/denominator;

   if (s < pgeneconversion)
   {
    geneConversion();
   }//end if

   else if (s < pretrotransposition)
   {
    retrotransposition();
   }//end else

   else
   {
    mutation();
   }//end else
}
//-----------------------------------------------------------------
double Subfamily::theta()

{
   double thetaSum = 0;
double thetaSumReal;
   double thetaDenominator = 0;
   double theta = 0;
double thetaReal;
int variableSitesReal;
double thetaDenominatorReal;
   double j;
double k;

 for (double i = 1; i <= 72; i++)
   {
      j = 1/i;
      thetaSum = thetaSum + j;
   }

   thetaDenominator = sequenceLength * thetaSum;

   theta = variableSites/thetaDenominator;

variableSitesReal = 190;

for (double i = 1; i <= elementsCreated; i++)
{
   k = 1/i;
   thetaSumReal = thetaSumReal + k;
}
thetaDenominatorReal = sequenceLength * thetaSumReal;
thetaReal = variableSitesReal/thetaDenominatorReal;

   return theta;
}
//-----------------------------------------------------------------
double Subfamily::pi()
{
   double pi = 0;
   double combinations = 0;
   nucleotideDifferences = 0;

   for (int n = 0; n < elements.size(); n++)
   {
      Element& elementQuery = elements.at(n);

      for (int i=0; i<elements.size(); i++)
      {
         Element& elementSubject = elements.at(i);
         if (n==i) continue;
         combinations++;

         for (int k=0; k<sequenceLength; k++)
         {
            char c1 = elementSubject.sequence.at(k);
            char c2 = elementQuery.sequence.at(k);

            if (c1 != c2)
            {
               nucleotideDifferences++;
            }
         }
      }//end for
   }
   nucleotideDifferences = nucleotideDifferences/2;
   combinations = combinations/2;
   meanNucleotideDifferences = nucleotideDifferences/combinations;
   pi = meanNucleotideDifferences/(double)sequenceLength;

   return pi;
}
//-------------------------------------------------------------------
//Main
int main()

{
   for (activeThreshold = 0.1; activeThreshold <= 1; activeThreshold += 0.1)
   {
      for (retroThreshold = 0.00001; retroThreshold <= 0.0001; retroThreshold += 0.00001)
      {

         try {
            double runs = 0;
            double averageSourceCount = 0;
            double averageTime = 0;
            double averageMutations = 0;
            double averageParallelMutations = 0;
            double averageSharedSourceMutations = 0;
            double averageParallelSites = 0;
            double averageVariableSites = 0;
            double averageTheta = 0;
            double averagePi = 0;
            double successes = 0;
            double percentSuccess = 0;

            for (runs = 0; runs <100; runs++)
            {
               //reset values for new run
               sourceCount = 0;
               parallelMutations = 0;
               sharedSourceMutations = 0;
               parallelSites = 0;
               thetaResult = 0;
               piResult = 0;

               a = 1;
               NTV = 0;
               NTS = 0;
               CTV = 0;
               CTS = 0;

               actives = 1;
               totalTime = 0;

                try {
               Subfamily subfamily;

               }
               catch (exception e) {
               exit(0);
                }

               //Test each run for success
               if (totalTime >= 10000000)
               {
                  if (totalTime <= 16000000)
                  {
                    if (thetaResult <= 0.10611128)
                     {
                        if (thetaResult >= 0.08681832)
                        {
                           if (piResult <= 1)
                           {
                              if (piResult >=0)
                              {
                              successes++;
                              cout << "   sources = " << sourceCount;
                              cout << "   shared mutations = " << sharedSourceMutations + parallelMutations;
                              cout << "   theta = " << thetaResult;
                              cout << "   pi = " << piResult;
                              cout << "   time = " << totalTime << "\n";
                              //cout << " pT" << retroThreshold << "\n";
                              }
                           }
                        }
                     }
                  }
               }

               averagePi = averagePi + piResult;
               averageTheta = averageTheta + thetaResult;
               averageTime = averageTime + totalTime;
               averageMutations = averageMutations + (NTV + NTS + CTS + CTV);
               averageParallelMutations = averageParallelMutations + parallelMutations;
               averageParallelSites = averageParallelSites + parallelSites;
               averageVariableSites = averageVariableSites + variableSites;
               averageSharedSourceMutations = averageSharedSourceMutations + sharedSourceMutations;
               averageSourceCount = averageSourceCount + sourceCount;

            }//end for - runs loop

             percentSuccess = successes/runs;
             percentSuccess = percentSuccess * 100;

            averagePi = averagePi/runs;
            averageTheta = averageTheta/runs;
            averageSourceCount = averageSourceCount/runs;
            averageTime = averageTime/runs;
            averageMutations = averageMutations/runs;
            averageParallelMutations = averageParallelMutations/runs;
            averageParallelSites = averageParallelSites/runs;
            averageVariableSites = averageVariableSites/runs;
            averageSharedSourceMutations = averageSharedSourceMutations/runs;

if (percentSuccess > 0)
{
cout << "Activation threshold: " << activeThreshold << "\n";
cout << "Retrotransposition threshold: " << retroThreshold << "\n";
cout << "Percent success: " << percentSuccess << "%" << "\n";
cout << "Average shared mutations: " << averageSharedSourceMutations + averageParallelMutations <<
"\n";
cout << "Average source count: " << averageSourceCount << "\n";
cout << "Average theta: " << averageTheta << "\n";
cout << "Average pi: " << averagePi << "\n";
cout << "Average total time: " << averageTime << "\n";
}
            if (averageTime >= 10000000)
            {
               if (averageTime <= 16000000)
               {
                 if (averageTheta <= 0.10611128)
                  {
                     if (averageTheta >=0.08681832)
                     {
                       cout << "Activation threshold: " << activeThreshold << "\n";
                       cout << "Retrotransposition threshold: " << retroThreshold << "\n";
                       cout << "Mutation threshold: " << mutationThreshold << "\n\n";
                       cout << "Average source count: " << averageSourceCount << "\n";
                       cout << "Average total mutations: " << averageMutations << "\n";
```

```
                cout << "Average total time: " << averageTime << "\n";
                cout << "Average shared mutations: " << averageSharedSourceMutations +
averageParallelMutations << "\n";
               cout << "Average shared source mutations: " << averageSharedSourceMutations << "\n";
                cout << "Average variable sites: " << averageVariableSites << "\n";
                cout << "Average theta: " << averageTheta << "\n";
                cout << "Average pi: " << averagePi << "\n";
                cout << "Percent success: " << percentSuccess << "%" << "\n";
                cout << "Average parallel mutations: " << averageParallelMutations << "\n\n";
                cout << "***********************************************************************\n\n";
              }
            }
          }
        }

      }//end try

      catch (int errorNum)
      {
        if (errorNum == 0)
        {
          cout << "No more CpGs, cannot generate subfamily" << "\n";
          continue;
        }
      }//end catch

    }//end for - retroThreshold

    cout <<
"~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~\n";
    cout <<
"~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~\n\n";
  }
}
```

File 2: Subfamily.h

```
#ifndef SUBFAMILY_H
#define SUBFAMILY_H

#include <iostream>
#include <string>
#include <vector>
#include <fstream>


using namespace std;

class Element;

class Subfamily
{
   public:
      vector<Element> elements;

      Subfamily();
      void mutation();
      double exponential();
      void eventDecision();
      double theta();
      double pi();
      void retrotransposition();
      void geneConversion();
      void gcCompleteForward();
      void gcCompleteBackward(int);
      void gcPartial();
      void gcPartialAluSg();
      void gcPartialAluY();
      int elementsCreated;
      int sequenceLength;
      double nucleotideDifferences;
      double meanNucleotideDifferences;




};


#endif // SUBFAMILY_H
```

File 3: Element.h

```
#ifndef ELEMENT_H
#define ELEMENT_H

#include <iostream>
#include <string>
#include <vector>

using namespace std;

extern int a;
extern int NTV;
extern int NTS;
extern int CTS;
extern int CTV;
extern int parallelMutations;
extern int parallelSites;


class Element
{
   public:
      Element() {}

      Element (Element*);

      Element* parentPtr;
      int designation;
      bool source;
      bool active;
      int parent;
      int position;
      string sequence;
      int numberOfMutations;
      void printMe();
      void Mutation();
      void positionSelect();
      void CpGtransversion();
      void CpGtransversion1();
      void CpGtransversion2();
      void CpGtransition();
      void transition();
      void transversion();
      void transversion1();
      void transversion2();

};



#endif // ELEMENT_H
```

File 4: Element.cpp

```
#include "Element.h"
#include "Subfamily.h"

#include <iostream>
#include <string>

using namespace std;

extern bool Activate();
extern void printMe();

extern bool parallelCTS[280];
extern bool parallelTS[280];
extern bool parallelTV1[280];
extern bool parallelTV2[280];
extern bool parallelCTV1[280];
extern bool parallelCTV2[280];
extern bool mutatedSites[280];

string YNactive = "error";
string YNsource = "error";

Element::Element(Element* parentPtr)
{
   designation = a;
   source = false;
   active = Activate();
   numberOfMutations = parentPtr->numberOfMutations;
   parent = parentPtr->designation;
   sequence = parentPtr->sequence;
}//end constructor

//-----------------------------------------------------------------
void Element::printMe()
{
   if (active == false)
      YNactive = "No";

   else
      YNactive = "Yes";


   if (source == false)
      YNsource = "No";
   else
      YNsource = "Yes";

cout << "Designation: " << designation << "\n"
       << "Number of mutations: " << numberOfMutations << "\n\n";
       << "Parent: " << parent << "\n"
       << "Sequence: " << sequence << "\n";
cout << ">" << designation << "\n" << sequence << "\n";
       << "Active? " << YNactive << "\n"
       << "Source? " << YNsource << "\n\n\n";

}

//-----------------------------------------------------------------
//  MUTATION FUNCTIONS

//-----------------------------------------------------------------
void Element::Mutation()
{
   position = 0;

   int typeNumber = rand() % 48;

   if (typeNumber <= 1)
   {
      CpGtransversion();
      CTV++;
   }
```

```
  else if (typeNumber <= 10)
  {  transversion();
    NTV++;
  }

  else if (typeNumber <=26)
  {
    transition();
    NTS++;
  }

  else
  {
    CpGtransition();
    CTS++;
  }

}

void Element::CpGtransversion()
{

  int t = rand() % 2;
  if (t == 0)
    CpGtransversion1();
  else
    CpGtransversion2();

}

//-----------------------------------------------------------------
void Element::transversion()
{
  int t = rand() % 2;
  if (t == 0)
    transversion1();
  else
    transversion2();
}

//-----------------------------------------------------------------
void Element::transition()
{

  bool done = false;

  //check that position is suitable for mutation (not a cpg site)
  while (!done)
  {
    positionSelect();

    if (sequence[position] == 'C')
    {
      //if rightmost base, fine to mutate
      if (position == sequence.size()-1) done = true;

      //if not, check +1 is not a G
      else if (sequence[position+1]=='G') continue;
      else done = true;
    }//end if

    else if (sequence[position] == 'G')
    {
      //if leftmost base, fine to mutate
      if (position != 0) break;

      //if not, check -1 is not a C
      if (sequence[position-1]=='C') continue;
      else done = true;
    }//end if

    else done = true;
  }//end while

    if (parallelTS[position] == true)
  {
    parallelMutations++;
  }
  else
  {
   parallelTS[position] = true;
  }

  //make mutation
  switch (sequence[position])
  {
  case 'A':
    sequence[position] = 'G';
    break;

  case 'G':
    sequence[position] = 'A';
    break;

  case 'T':
    sequence[position] = 'C';
    break;

  case 'C':
    sequence[position] = 'T';
    break;
  }//end switch
mutatedSites[position] = true;
}//end function


//-----------------------------------------------------------------
void Element::transversion1()
{
  bool done = false;

  //check that position is suitable for mutation (not a cpg site)
  while (!done)
  {
    positionSelect();

    if (sequence[position] == 'C')
    {
      //if rightmost base, fine to mutate
      if (position == sequence.size()-1) done = true;

      //if not, check +1 is not a G
      else if (sequence[position+1]=='G') continue;
      else done = true;
    }//end if

    else if (sequence[position] == 'G')
    {
      //if leftmost base, fine to mutate
      if (position != 0) break;

      //if not, check -1 is not a C
      if (sequence[position-1]=='C') continue;
      else done = true;
    }//end if

    else done = true;
  }//end while

    if (parallelTV1[position] == true)
  {
    parallelMutations++;
  }
  else
  {
   parallelTV1[position] = true;
  }

  //make mutation
  switch (sequence[position])
  {
  case 'A':
    sequence[position] = 'T';
    break;

  case 'G':
    sequence[position] = 'C';
    break;

  case 'T':
    sequence[position] = 'G';
    break;

  case 'C':
    sequence[position] = 'A';
    break;
}
mutatedSites[position] = true;
}

//-----------------------------------------------------------------
void Element::transversion2()
{
  bool done = false;

  //check that position is suitable for mutation (not a cpg site)
  while (!done)
  {
    positionSelect();

    if (sequence[position] == 'C')
    {
      //if rightmost base, fine to mutate
      if (position == sequence.size()-1) done = true;

      //if not, check +1 is not a G
      else if (sequence[position+1]=='G') continue;
      else done = true;
    }//end if

    else if (sequence[position] == 'G')
    {
      //if leftmost base, fine to mutate
      if (position != 0) break;

      //if not, check -1 is not a C
      if (sequence[position-1]=='C') continue;
      else done = true;
    }//end if

    else done = true;
  }//end while

    if (parallelTV2[position] == true)
  {
    parallelMutations++;
  }
  else
  {
   parallelTV2[position] = true;
  }

  //make mutation
  switch (sequence[position])
  {
  case 'A':
    sequence[position] = 'C';
    break;

  case 'G':
    sequence[position] = 'T';
    break;

  case 'T':
    sequence[position] = 'A';
```

```
      break;

   case 'C':
      sequence[position] = 'G';
      break;
   }
mutatedSites[position] = true;
}

//-----------------------------------------------------------------
void Element::CpGtransition()
{

   if (sequence.find("CG") == -1)
   {
   throw(0);

   }
   bool done = false;

   //check that position is suitable for mutation (is a CpG site)
   while (!done)
   {
      positionSelect();

      if (sequence[position] == 'C')
      {
         //if rightmost base, try again
         if (position == sequence.size()-1) continue;

         //if not, ensure pos+1 is a G
         if (sequence[position+1]=='G') done = true;
      }

      else if (sequence[position] == 'G')
      {
         //if leftmost base, try again
         if (position != 0) continue;

         //if not, ensure pos-1 is a C
         if (sequence[position-1]=='C') continue;
      }//end else

      //if position is not C or G, try again
      else positionSelect();
   }

   if (parallelCTS[position] == true)
   {
      parallelMutations++;
   }
   else
   {
    parallelCTS[position] = true;
   }
   //make mutation
   switch (sequence[position])
   {
   case 'G':
      sequence[position] = 'A';
      break;
   case 'C':
      sequence[position] = 'T';
      break;
   }
mutatedSites[position] = true;
}

//-----------------------------------------------------------------

void Element::CpGtransversion1()
{

   if (sequence.find("CG") == -1) throw(0);
```

```
  bool done = false;

  //check that position is suitable for mutation (is a CpG site)
  while (!done)
  {
    positionSelect();

    if (sequence[position] == 'C')
    {
      //if rightmost base, try again
      if (position == sequence.size()-1) continue;

      //if not, ensure pos+1 is a G
      if (sequence[position+1]=='G') done = true;
    }

    else if (sequence[position] == 'G')
    {
      //if leftmost base, try again
      if (position != 0) continue;

      //if not, ensure pos-1 is a C
      if (sequence[position-1]=='C') continue;
    }//end else

    //if position is not C or G, try again
    else positionSelect();
  }

    if (parallelCTV1[position] == true)
  {
    parallelMutations++;
  }
  else
  {
   parallelCTV1[position] = true;
  }


  //make mutation
  switch (sequence[position])
  {
  case 'G':
    sequence[position] = 'C';
    break;
  case 'C':
    sequence[position] = 'A';
    break;

  }
mutatedSites[position] = true;
}

//-----------------------------------------------------------------
void Element::CpGtransversion2()
{
  if (sequence.find("CG") == -1) throw(0);

  bool done = false;

  //check that position is suitable for mutation (is a CpG site)
  while (!done)
  {
    positionSelect();

    if (sequence[position] == 'C')
    {
      //if rightmost base, try again
      if (position == sequence.size()-1) continue;

      //if not, ensure pos+1 is a G
      if (sequence[position+1]=='G') done = true;
    }

    else if (sequence[position] == 'G')
    {
      //if leftmost base, try again
      if (position != 0) continue;

      //if not, ensure pos-1 is a C
      if (sequence[position-1]=='C') continue;
    }//end else

    //if position is not C or G, try again
    else positionSelect();
  }


    if (parallelCTV2[position] == true)
  {
    parallelMutations++;
  }
  else
  {
   parallelCTV2[position] = true;
  }
  //make mutation
  switch (sequence[position])
  {
  case 'G':
    sequence[position] = 'T';
    break;
  case 'C':
    sequence[position] = 'G';
    break;
  }//end switch

mutatedSites[position] = true;
}

//---------------------------------------------------------------
void Element::positionSelect()
{
  position = rand() % sequence.length();
}
```

Appendix 3 - Smallest divergence for all interspecies comparisons of DNA transposons in *Drosophila*

In appendices 3, 4 and 5, yellow shading is used to indicate interspecies comparisions of transposable elements within a particular family for which the smallest divergence observed is smaller than that between the host *Adh* coding regions. Green shading indicates that the smallest divergence slightly exceeded that between the *Adh* coding regions. White shading indicates divergence between transposable elements in excess of that between the *Adh* coding regions. Where a particular transposable element family is not observed in both species 1 and species 2 shown in the leftmost columns, the cell is shaded in grey to indicate that no comparison was possible.

species 1	species 2	Adh	Bari	DNAREP1	hAT1	hAT1N	Helitron1	Helitron1_Dvir	Hobo	Looper	M4
mel	sim	0.019	0.001	0.133					0.001	0.036	0.065
mel	sec	0.017	0.003	0.154					0.001	0.045	0.049
mel	yak	0.044		0.132					0.174	0.125	
mel	ere	0.049	0.298	0.165					0.253		
mel	ana	0.110	0.165				0.058			0.244	
mel	pse	0.142									
mel	per	0.139									
mel	vir	0.237								0.010	
mel	moj	0.220									
mel	gri	0.246									
sim	sec	0.005	0.003	0.014					0.000	0.015	0.017
sim	yak	0.038		0.129					0.271	0.133	
sim	ere	0.045	0.299	0.161					0.282		
sim	ana	0.101	0.166							0.245	
sim	pse	0.145									
sim	per	0.141									
sim	vir	0.231								0.014	
sim	moj	0.216									
sim	gri	0.239									
sec	yak	0.035		0.123					0.130	0.141	
sec	ere	0.045	0.293	0.154					0.258		
sec	ana	0.100	0.166							0.239	
sec	pse	0.144									
sec	per	0.140									
sec	vir	0.231								0.029	
sec	moj	0.216									
sec	gri	0.239									
yak	ere	0.051		0.115					0.281		
yak	ana	0.100								0.307	
yak	pse	0.141									
yak	per	0.137									
yak	vir	0.237								0.150	
yak	moj	0.220									
yak	gri	0.242									
ere	ana	0.102	0.240								
ere	pse	0.136									
ere	per	0.135									
ere	vir	0.231									
ere	moj	0.220									
ere	gri	0.235									
ana	pse	0.159				0.323					
ana	per	0.158				0.274					
ana	vir	0.252								0.215	
ana	moj	0.226				0.395					
ana	gri	0.252									
pse	per	0.004			0.007	0.005					
pse	vir	0.224									
pse	moj	0.205			0.271	0.145					
pse	gri	0.226									
per	vir	0.221									
per	moj	0.207			0.280	0.087					
per	gri	0.227									
vir	moj	0.144						0.147			
vir	gri	0.176									
moj	gri	0.176									

species 1	species 2	Adh	Marina	Mariner	Mariner2	Marw olen	Marw olen2	Minos	NOF FB	Paris
mel	sim	0.019			0.059					
mel	sec	0.017			0.050					
mel	yak	0.044								
mel	ere	0.049								
mel	ana	0.110								
mel	pse	0.142								
mel	per	0.139								
mel	vir	0.237								
mel	moj	0.220								
mel	gri	0.246								
sim	sec	0.005		0.007	0.014					
sim	yak	0.038		0.012						
sim	ere	0.045								
sim	ana	0.101								
sim	pse	0.145								
sim	per	0.141								
sim	vir	0.231								
sim	moj	0.216								
sim	gri	0.239								
sec	yak	0.035		0.013						
sec	ere	0.045								
sec	ana	0.100								
sec	pse	0.144								
sec	per	0.140								
sec	vir	0.231								
sec	moj	0.216								
sec	gri	0.239								
yak	ere	0.051							0.086	
yak	ana	0.100	0.323							
yak	pse	0.141								
yak	per	0.137								
yak	vir	0.237								
yak	moj	0.220					0.324	0.140		
yak	gri	0.242								
ere	ana	0.102								
ere	pse	0.136								
ere	per	0.135								
ere	vir	0.231								
ere	moj	0.220								
ere	gri	0.235								
ana	pse	0.159				0.193				0.291
ana	per	0.158				0.239				0.248
ana	vir	0.252								0.273
ana	moj	0.226								0.302
ana	gri	0.252								
pse	per	0.004				0.041				0.010
pse	vir	0.224								0.303
pse	moj	0.205								0.206
pse	gri	0.226								
per	vir	0.221								0.262
per	moj	0.207								0.180
per	gri	0.227								
vir	moj	0.144								0.308
vir	gri	0.176								
moj	gri	0.176								

species 1	species 2	Adh	ProtoP	ProtoP A	ProtoP B	Rehavkus1	S	S2	Tc1	Tc1-2
mel	sim	0.019	0.110	0.133	0.087		0.095		0.081	0.076
mel	sec	0.017	0.037	0.141	0.081		0.101		0.064	0.065
mel	yak	0.044	0.105		0.209				0.090	0.103
mel	ere	0.049	0.102						0.069	0.248
mel	ana	0.110								0.384
mel	pse	0.142						0.256		
mel	per	0.139						0.259		
mel	vir	0.237						0.331		
mel	moj	0.220						0.321		
mel	gri	0.246								
sim	sec	0.005	0.028	0.043	0.032		0.054		0.016	0.019
sim	yak	0.038	0.153		0.207				0.102	0.122
sim	ere	0.045	0.116						0.084	0.227
sim	ana	0.101								0.349
sim	pse	0.145								
sim	per	0.141								
sim	vir	0.231								
sim	moj	0.216								
sim	gri	0.239								
sec	yak	0.035	0.073		0.205				0.085	0.120
sec	ere	0.045	0.098						0.066	0.222
sec	ana	0.100								0.354
sec	pse	0.144								
sec	per	0.140								
sec	vir	0.231								
sec	moj	0.216								
sec	gri	0.239								
yak	ere	0.051	0.089						0.067	0.196
yak	ana	0.100				0.373				0.418
yak	pse	0.141								
yak	per	0.137								
yak	vir	0.237								
yak	moj	0.220								
yak	gri	0.242								
ere	ana	0.102								0.411
ere	pse	0.136								
ere	per	0.135								
ere	vir	0.231								
ere	moj	0.220								
ere	gri	0.235								
ana	pse	0.159								
ana	per	0.158								
ana	vir	0.252								
ana	moj	0.226								
ana	gri	0.252								
pse	per	0.004						0.006		
pse	vir	0.224						0.261		
pse	moj	0.205						0.270		
pse	gri	0.226								
per	vir	0.221						0.217		
per	moj	0.207						0.234		
per	gri	0.227								
vir	moj	0.144						0.276		
vir	gri	0.176								
moj	gri	0.176								

species 1	species 2	Adh	Transib1	Transib1_DP	Transib2	Transib2_DP	Transib3	Transib3_DP	Transib4
mel	sim	0.019	0.042		0.029		0.100		0.199
mel	sec	0.017	0.028		0.010		0.083		0.099
mel	yak	0.044	0.111		0.116		0.290		0.156
mel	ere	0.049			0.057				
mel	ana	0.110	0.390						
mel	pse	0.142							
mel	per	0.139							
mel	vir	0.237							
mel	moj	0.220	0.238						
mel	gri	0.246							
sim	sec	0.005	0.011		0.015		0.020		0.030
sim	yak	0.038	0.088		0.137		0.218		0.245
sim	ere	0.045			0.067				
sim	ana	0.101	0.357						
sim	pse	0.145							
sim	per	0.141							
sim	vir	0.231							
sim	moj	0.216	0.287d						
sim	gri	0.239							
sec	yak	0.035	0.088d		0.042		0.118		0.427
sec	ere	0.045			0.052				
sec	ana	0.100	0.328						
sec	pse	0.144							
sec	per	0.140							
sec	vir	0.231							
sec	moj	0.216	0.284						
sec	gri	0.239							
yak	ere	0.051			0.087				
yak	ana	0.100	0.428						
yak	pse	0.141							
yak	per	0.137							
yak	vir	0.237							
yak	moj	0.220	0.415						
yak	gri	0.242							
ere	ana	0.102							
ere	pse	0.136							
ere	per	0.135							
ere	vir	0.231							
ere	moj	0.220							
ere	gri	0.235							
ana	pse	0.159		0.196					
ana	per	0.158		0.191					
ana	vir	0.252							
ana	moj	0.226	0.353	0.298					
ana	gri	0.252							
pse	per	0.004		0.016		0.005		0.011	
pse	vir	0.224							
pse	moj	0.205		0.313					
pse	gri	0.226							
per	vir	0.221							
per	moj	0.207		0.318					
per	gri	0.227							
vir	moj	0.144							
vir	gri	0.176							
moj	gri	0.176							

species 1	species 2	Adh	Transib N2	Transib N3	Transib N4	Transib N5	UHU
mel	sim	0.019					
mel	sec	0.017					
mel	yak	0.044					
mel	ere	0.049					
mel	ana	0.110					
mel	pse	0.142					
mel	per	0.139					
mel	vir	0.237					
mel	moj	0.220					
mel	gri	0.246					
sim	sec	0.005					
sim	yak	0.038					
sim	ere	0.045					
sim	ana	0.101					
sim	pse	0.145					
sim	per	0.141					
sim	vir	0.231					
sim	moj	0.216					
sim	gri	0.239					
sec	yak	0.035					
sec	ere	0.045					
sec	ana	0.100					
sec	pse	0.144					
sec	per	0.140					
sec	vir	0.231					
sec	moj	0.216					
sec	gri	0.239					
yak	ere	0.051					
yak	ana	0.100					
yak	pse	0.141					
yak	per	0.137					
yak	vir	0.237					
yak	moj	0.220					
yak	gri	0.242					
ere	ana	0.102					
ere	pse	0.136					
ere	per	0.135					
ere	vir	0.231					
ere	moj	0.220					
ere	gri	0.235					
ana	pse	0.159					
ana	per	0.158					
ana	vir	0.252					
ana	moj	0.226					
ana	gri	0.252					0.395
pse	per	0.004	0.000	0.166	0.038	0.007	
pse	vir	0.224					
pse	moj	0.205					
pse	gri	0.226					
per	vir	0.221					
per	moj	0.207					
per	gri	0.227					
vir	moj	0.144					
vir	gri	0.176					
moj	gri	0.176					

Appendix 4 - Smallest divergence for all interspecies comparisons of non-LTR retrotransposons in *Drosophila*

species 1	species 2	Adh	Alu II	Baggins1	Bilbo	BS	BS2	BS3	BS4	CR1A
mel	sim	0.019	0.082	0.034		0.049	0.007	0.086	0.043	0.043
mel	sec	0.017	0.183	0.033		0.043	0.023	0.089	0.055	0.045
mel	yak	0.044					0.046	0.229		0.135
mel	ere	0.049				0.182	0.104	0.248		0.151
mel	ana	0.110								
mel	pse	0.142								
mel	per	0.139				0.208				
mel	vir	0.237								
mel	moj	0.220				0.241	0.243			
mel	gri	0.246								
sim	sec	0.005	0.027	0.046		0.017	0.025	0.028	0.019	0.021
sim	yak	0.038					0.026	0.163		0.131
sim	ere	0.045				0.199	0.098	0.181		0.151
sim	ana	0.101								
sim	pse	0.145								
sim	per	0.141				0.233				
sim	vir	0.231								
sim	moj	0.216				0.261	0.240			
sim	gri	0.239								
sec	yak	0.035					0.046	0.268		0.134
sec	ere	0.045				0.178	0.101	0.279		0.161
sec	ana	0.100								
sec	pse	0.144								
sec	per	0.140				0.232				
sec	vir	0.231								
sec	moj	0.216				0.255	0.244			
sec	gri	0.239								
yak	ere	0.051					0.096	0.093		0.149
yak	ana	0.100								
yak	pse	0.141								
yak	per	0.137								
yak	vir	0.237								
yak	moj	0.220					0.229			
yak	gri	0.242								
ere	ana	0.102								
ere	pse	0.136								
ere	per	0.135				0.211				
ere	vir	0.231								
ere	moj	0.220				0.253	0.258			
ere	gri	0.235								
ana	pse	0.159								
ana	per	0.158								
ana	vir	0.252								
ana	moj	0.226								
ana	gri	0.252								
pse	per	0.004			0.187					
pse	vir	0.224								
pse	moj	0.205								
pse	gri	0.226								
per	vir	0.221								
per	moj	0.207				0.213				
per	gri	0.227								
vir	moj	0.144								
vir	gri	0.176								
moj	gri	0.176								

species 1	species 2	Adh	doc	doc2	doc3	doc4	doc5	doc6	FW	FW2
mel	sim	0.019	0.001	0.018	0.084	0.101	0.074	0.001	0.023	0.074
mel	sec	0.017	0.002	0.017	0.058	0.132	0.062	0.001	0.024	0.222
mel	yak	0.044	0.007	0.081	0.109			0.001	0.051	0.296
mel	ere	0.049		0.180					0.190	
mel	ana	0.110								
mel	pse	0.142								
mel	per	0.139								
mel	vir	0.237								
mel	moj	0.220								
mel	gri	0.246								
sim	sec	0.005	0.003	0.019	0.025	0.044	0.020	0.000	0.001	0.020
sim	yak	0.038	0.006	0.091	0.137			0.001	0.051	0.141
sim	ere	0.045		0.182					0.191	
sim	ana	0.101								
sim	pse	0.145								
sim	per	0.141								
sim	vir	0.231								
sim	moj	0.216								
sim	gri	0.239								
sec	yak	0.035	0.009	0.094	0.119			0.002	0.049	0.092
sec	ere	0.045		0.183					0.189	
sec	ana	0.100								
sec	pse	0.144								
sec	per	0.140								
sec	vir	0.231								
sec	moj	0.216								
sec	gri	0.239								
yak	ere	0.051		0.211					0.172	
yak	ana	0.100								
yak	pse	0.141								
yak	per	0.137								
yak	vir	0.237								
yak	moj	0.220								
yak	gri	0.242								
ere	ana	0.102								
ere	pse	0.136								
ere	per	0.135								
ere	vir	0.231								
ere	moj	0.220								
ere	gri	0.235								
ana	pse	0.159								
ana	per	0.158								
ana	vir	0.252								
ana	moj	0.226								
ana	gri	0.252								
pse	per	0.004								
pse	vir	0.224								
pse	moj	0.205								
pse	gri	0.226								
per	vir	0.221								
per	moj	0.207								
per	gri	0.227								
vir	moj	0.144								
vir	gri	0.176								
moj	gri	0.176								

species 1	species 2	Adh	G	G2	G4	G5	G5A	G6	Helena_DS	HetA
mel	sim	0.019	0.420	0.057	0.085	0.056	0.070	0.005	0.076	0.112
mel	sec	0.017	0.156	0.055	0.095	0.052	0.073	0.001	0.077	0.119
mel	yak	0.044	0.069	0.123	0.182				0.160	
mel	ere	0.049	0.274	0.084	0.081		0.101		0.181	
mel	ana	0.110							0.267	
mel	pse	0.142								
mel	per	0.139								
mel	vir	0.237							0.264	
mel	moj	0.220							0.271	
mel	gri	0.246								
sim	sec	0.005	0.088	0.025	0.034	0.011	0.015	0.003	0.011	0.001
sim	yak	0.038	0.240	0.089	0.642				0.103	
sim	ere	0.045	0.487	0.102	0.080		0.115		0.177	
sim	ana	0.101							0.248	
sim	pse	0.145								
sim	per	0.141								
sim	vir	0.231							0.256	
sim	moj	0.216							0.235	
sim	gri	0.239								
sec	yak	0.035	0.151	0.041	0.154				0.108	
sec	ere	0.045	0.260	0.075	0.066		0.165		0.110	
sec	ana	0.100							0.238	
sec	pse	0.144								
sec	per	0.140								
sec	vir	0.231							0.248	
sec	moj	0.216							0.229	
sec	gri	0.239								
yak	ere	0.051	0.253	0.106	0.121				0.191	
yak	ana	0.100							0.261	
yak	pse	0.141								
yak	per	0.137								
yak	vir	0.237							0.257	
yak	moj	0.220							0.240	
yak	gri	0.242								
ere	ana	0.102							0.281	
ere	pse	0.136								
ere	per	0.135								
ere	vir	0.231							0.302	
ere	moj	0.220							0.267	
ere	gri	0.235								
ana	pse	0.159								
ana	per	0.158								
ana	vir	0.252							0.235	
ana	moj	0.226							0.239	
ana	gri	0.252								
pse	per	0.004								
pse	vir	0.224								
pse	moj	0.205								
pse	gri	0.226								
per	vir	0.221								
per	moj	0.207								
per	gri	0.227								
vir	moj	0.144							0.187	
vir	gri	0.176								
moj	gri	0.176								

species 1	species 2	Adh	IVK	Jockey2	LINE J-1	R1	R2_DM	RT1A	RT1B	RT1C
mel	sim	0.019	0.026	0.065	0.003	0.043	0.032	0.597	0.032	0.065
mel	sec	0.017	0.040	0.065	0.006	0.049	0.022	0.107	0.041	0.069
mel	yak	0.044	0.144	0.200		0.086	0.094	0.194	0.251	0.251
mel	ere	0.049		0.109		0.118		0.245	0.257	0.277
mel	ana	0.110			0.253		0.007			
mel	pse	0.142								
mel	per	0.139								
mel	vir	0.237								
mel	moj	0.220								
mel	gri	0.246								
sim	sec	0.005	0.020	0.025	0.003	0.019	0.007	0.618	0.034	0.042
sim	yak	0.038	0.126	0.221		0.136	0.107	0.617	0.273	0.206
sim	ere	0.045		0.118		0.117		0.639	0.289	0.235
sim	ana	0.101			0.237		0.032			
sim	pse	0.145								
sim	per	0.141								
sim	vir	0.231								
sim	moj	0.216								
sim	gri	0.239								
sec	yak	0.035	0.150	0.172		0.118	0.101	0.253	0.251	0.208
sec	ere	0.045		0.107		0.122		0.372	0.233	0.243
sec	ana	0.100			0.227		0.035			
sec	pse	0.144								
sec	per	0.140								
sec	vir	0.231								
sec	moj	0.216								
sec	gri	0.239								
yak	ere	0.051		0.185		0.098		0.238	0.316	0.280
yak	ana	0.100					0.102			
yak	pse	0.141								
yak	per	0.137								
yak	vir	0.237								
yak	moj	0.220								
yak	gri	0.242								
ere	ana	0.102								
ere	pse	0.136								
ere	per	0.135								
ere	vir	0.231								
ere	moj	0.220								
ere	gri	0.235								
ana	pse	0.159								
ana	per	0.158								
ana	vir	0.252								
ana	moj	0.226								
ana	gri	0.252								
pse	per	0.004								
pse	vir	0.224								
pse	moj	0.205								
pse	gri	0.226								
per	vir	0.221								
per	moj	0.207								
per	gri	0.227								
vir	moj	0.144								
vir	gri	0.176								
moj	gri	0.176								

species 1	species 2	Adh	Spock	Tahre	Tart	TLD1	TLD2	TLD3	Trim	Worf
mel	sim	0.019		0.142	0.103		0.037	0.006		
mel	sec	0.017		0.143	0.063		0.037	0.000		
mel	yak	0.044			0.064	0.000	0.057			
mel	ere	0.049		0.266	0.182	0.000				
mel	ana	0.110								
mel	pse	0.142								
mel	per	0.139								
mel	vir	0.237								
mel	moj	0.220								
mel	gri	0.246								
sim	sec	0.005		0.033	0.006		0.000	0.006		
sim	yak	0.038			0.629		0.058			
sim	ere	0.045		0.263	0.175					
sim	ana	0.101								
sim	pse	0.145								
sim	per	0.141								
sim	vir	0.231								
sim	moj	0.216								
sim	gri	0.239								
sec	yak	0.035			0.621		0.052			
sec	ere	0.045		0.258	0.161					
sec	ana	0.100								
sec	pse	0.144								
sec	per	0.140								
sec	vir	0.231								
sec	moj	0.216								
sec	gri	0.239								
yak	ere	0.051			0.626	0.000				
yak	ana	0.100								
yak	pse	0.141								
yak	per	0.137								
yak	vir	0.237								
yak	moj	0.220								
yak	gri	0.242								
ere	ana	0.102								
ere	pse	0.136								
ere	per	0.135								
ere	vir	0.231								
ere	moj	0.220								
ere	gri	0.235								
ana	pse	0.159								
ana	per	0.158								
ana	vir	0.252								
ana	moj	0.226								
ana	gri	0.252								
pse	per	0.004	0.014						0.055	0.072
pse	vir	0.224								
pse	moj	0.205								
pse	gri	0.226								
per	vir	0.221								
per	moj	0.207								
per	gri	0.227								
vir	moj	0.144								
vir	gri	0.176								
moj	gri	0.176								

Appendix 5 - Smallest divergence for all interspecies comparisons of LTR retrotransposons in *Drosophila*

species 1	species 2	Adh	1731		297		412		Accord		Accord2		Batumi		Bel		Blastopia	
			I	LTR	I	LTR	I	LTR	I	LTR	I	LTR	I	LTR	I	LTR	I	LTR
mel	sim	0.019	0.003	0.008	0.004	0.000	0.001	0.015	0.001	0.000	0.049	0.009	0.007	0.026	0.003	0.003	0.001	0.004
mel	sec	0.017	0.007	0.007	0.011	0.000	0.009	0.029	0.000	0.000	0.025	0.009	0.021	0.024	0.008	0.011	0.002	0.011
mel	yak	0.044			0.135						0.070	0.206	0.025	0.033	0.006	0.009		
mel	ere	0.049			0.147						0.066	0.148						
mel	ana	0.110			0.217						0.099						0.357	0.455
mel	pse	0.142															0.353	0.519
mel	per	0.139															0.283	0.511
mel	wil	0.208																
mel	vir	0.237																
mel	moj	0.220																
mel	gri	0.246																
sim	sec	0.005	0.014	0.008	0.011	0.000	0.011	0.021	0.000	0.000	0.026	0.005	0.019	0.004	0.011	0.011	0.000	0.008
sim	yak	0.038			0.128						0.091	0.182	0.036	0.042	0.013	0.015		
sim	ere	0.045			0.110						0.091	0.133						
sim	ana	0.101			0.136						0.105						0.383	0.456
sim	pse	0.145															0.353	0.519
sim	per	0.141															0.288	0.511
sim	wil	0.203																
sim	vir	0.231																
sim	moj	0.216																
sim	gri	0.239																
sec	yak	0.035			0.126						0.064	0.177	0.032	0.037	0.011	0.012		
sec	ere	0.045			0.112						0.062	0.120						
sec	ana	0.100			0.134						0.087						0.302	0.451
sec	pse	0.144															0.324	0.516
sec	per	0.140															0.284	0.508
sec	wil	0.201																
sec	vir	0.231																
sec	moj	0.216																
sec	gri	0.239																
yak	ere	0.051			0.145						0.072	0.023						
yak	ana	0.100			0.147						0.097							
yak	pse	0.141																
yak	per	0.137																
yak	wil	0.204																
yak	vir	0.237																
yak	moj	0.220																
yak	gri	0.242																
ere	ana	0.102			0.173						0.097							
ere	pse	0.136																
ere	per	0.135																
ere	wil	0.204																
ere	vir	0.231																
ere	moj	0.220																
ere	gri	0.235																
ana	pse	0.159															0.410	0.592
ana	per	0.158															0.319	0.603
ana	wil	0.225																
ana	vir	0.252																
ana	moj	0.226																
ana	gri	0.252																
pse	per	0.004															0.014	0.000
pse	wil	0.174																
pse	vir	0.224																
pse	moj	0.205																
pse	gri	0.226																
per	wil	0.173																
per	vir	0.221																
per	moj	0.207																
per	gri	0.227																
wil	vir	0.238																
wil	moj	0.220																
wil	gri	0.231																
vir	moj	0.144																
vir	gri	0.176																
moj	gri	0.176																

species 1	species 2	Adh	Blood		Burdock		Chimpo		Circe	Copia		Copia2		Diver		Diver2		Frogger	
			I	LTR	I	LTR	I	LTR	I	I	LTR	I	LTR	I	LTR	I	LTR	I	LTR
mel	sim	0.019	0.005	0.002	0.004	0.004	0.006	0.015	0.039	0.005	0.022	0.047	0.083	0.008	0.009	0.039	0.031	0.031	0.025
mel	sec	0.017	0.009	0.029	0.014	0.019	0.006	0.015	0.025	0.005	0.051	0.028	0.088	0.003	0.018	0.036	0.023	0.029	0.020
mel	yak	0.044			0.022	0.035			0.036			0.317		0.001	0.000	0.042	0.036		
mel	ere	0.049			0.255				0.156										
mel	ana	0.110			0.168							0.337							
mel	pse	0.142																	
mel	per	0.139										0.333							
mel	wil	0.208								0.184									
mel	vir	0.237																	
mel	moj	0.220										0.404							
mel	gri	0.246										0.426							
sim	sec	0.005	0.003	0.018	0.013	0.007	0.003	0.011	0.021	0.004	0.021	0.028	0.006	0.002	0.013	0.031	0.005	0.031	0.034
sim	yak	0.038			0.018	0.035			0.045			0.312		0.005	0.009	0.037	0.037		
sim	ere	0.045			0.240				0.139										
sim	ana	0.101			0.166							0.351							
sim	pse	0.145																	
sim	per	0.141										0.366							
sim	wil	0.203								0.181									
sim	vir	0.231																	
sim	moj	0.216										0.400							
sim	gri	0.239										0.410							
sec	yak	0.035			0.018	0.041			0.037			0.321		0.006	0.018	0.037	0.033		
sec	ere	0.045			0.245				0.137										
sec	ana	0.100			0.133							0.358							
sec	pse	0.144																	
sec	per	0.140										0.325							
sec	wil	0.201								0.179									
sec	vir	0.231																	
sec	moj	0.216										0.406							
sec	gri	0.239										0.427							
yak	ere	0.051			0.235				0.130										
yak	ana	0.100			0.148							0.394							
yak	pse	0.141																	
yak	per	0.137										0.568							
yak	wil	0.204																	
yak	vir	0.237																	
yak	moj	0.220										0.446							
yak	gri	0.242										0.476							
ere	ana	0.102			0.219														
ere	pse	0.136																	
ere	per	0.135																	
ere	wil	0.204																	
ere	vir	0.231																	
ere	moj	0.220																	
ere	gri	0.235																	
ana	pse	0.159																	
ana	per	0.158										0.466							
ana	wil	0.225																	
ana	vir	0.252																	
ana	moj	0.226										0.382							
ana	gri	0.252										0.352							
pse	per	0.004																	
pse	wil	0.174																	
pse	vir	0.224																	
pse	moj	0.205																	
pse	gri	0.226																	
per	wil	0.173																	
per	vir	0.221																	
per	moj	0.207										0.466							
per	gri	0.227										0.480							
wil	vir	0.238																	
wil	moj	0.220																	
wil	gri	0.231																	
vir	moj	0.144																	
vir	gri	0.176																	
moj	gri	0.176										0.401							

species 1	species 2	Adh	Gtwin		Gypsy		ypsy_D	Gypsy2		Gypsy4		Gypsy5		Gypsy6		Gypsy7		Gypsy8	
			I	LTR	I	LTR	I	I	LTR	I	LTR	I	LTR	I	LTR	I	LTR	I	LTR
mel	sim	0.019	0.013	0.005	0.087	0.191		0.125		0.014	0.017	0.017	0.002			0.050	0.041	0.142	0.037
mel	sec	0.017	0.012	0.008	0.079	0.186		0.132		0.025	0.016					0.057	0.053		0.043
mel	yak	0.044	0.077	0.067	0.020	0.024		0.049	0.032	0.309		0.004	0.002	0.029		0.226		0.259	0.230
mel	ere	0.049	0.008	0.004	0.027	0.027		0.026	0.038	0.301		0.171	0.114	0.026	0.174			0.289	0.239
mel	ana	0.110	0.345							0.333				0.227	0.121				
mel	pse	0.142																	
mel	per	0.139																	
mel	wil	0.208																	
mel	vir	0.237												0.267					
mel	moj	0.220	0.276											0.445					
mel	gri	0.246																	
sim	sec	0.005	0.006	0.008	0.012	0.009		0.041		0.028	0.010					0.022	0.021		0.012
sim	yak	0.038	0.067	0.063	0.140	0.051		0.167		0.312		0.004	0.000			0.244		0.206	0.235
sim	ere	0.045	0.058	0.060	0.090	0.037		0.117		0.303		0.169	0.114					0.433	0.248
sim	ana	0.101	0.334							0.344									
sim	pse	0.145																	
sim	per	0.141																	
sim	wil	0.203																	
sim	vir	0.231																	
sim	moj	0.216	0.281																
sim	gri	0.239																	
sec	yak	0.035	0.063	0.118	0.132	0.038		0.153		0.314						0.374			0.235
sec	ere	0.045	0.048	0.089	0.078	0.023		0.128		0.306									0.251
sec	ana	0.100	0.326							0.303									
sec	pse	0.144																	
sec	per	0.140																	
sec	wil	0.201																	
sec	vir	0.231																	
sec	moj	0.216	0.246																
sec	gri	0.239																	
yak	ere	0.051	0.035	0.044	0.018	0.021		0.084	0.026	0.007		0.021	0.093	0.021				0.216	0.118
yak	ana	0.100	0.330							0.322				0.231					
yak	pse	0.141																	
yak	per	0.137																	
yak	wil	0.204																	
yak	vir	0.237												0.277					
yak	moj	0.220	0.283											0.432					
yak	gri	0.242																	
ere	ana	0.102	0.326							0.321				0.215	0.162				
ere	pse	0.136																	
ere	per	0.135																	
ere	wil	0.204																	
ere	vir	0.231												0.277					
ere	moj	0.220	0.243											0.311					
ere	gri	0.235																	
ana	pse	0.159																	
ana	per	0.158																	
ana	wil	0.225																	
ana	vir	0.252												0.271					
ana	moj	0.226	0.311											0.464					
ana	gri	0.252																	
pse	per	0.004																	
pse	wil	0.174																	
pse	vir	0.224																	
pse	moj	0.205																	
pse	gri	0.226																	
per	wil	0.173																	
per	vir	0.221					0.052												
per	moj	0.207																	
per	gri	0.227																	
wil	vir	0.238																	
wil	moj	0.220																	
wil	gri	0.231																	
vir	moj	0.144												0.047					
vir	gri	0.176																	
moj	gri	0.176																	

species 1	species 2	Adh	Gypsy9		Gypsy10		Gypsy12		IS_Beag	Idefix		Invader1		Invader2		Invader3		Invader4	
			I	LTR	I	LTR	I	LTR	I	I	LTR	I	LTR	I	LTR	I	LTR	I	LTR
mel	sim	0.019	0.113	0.036	0.075	0.000	0.019	0.054	0.011	0.168	0.102	0.008	0.007	0.039	0.022	0.017	0.013	0.129	0.074
mel	sec	0.017	0.103	0.028	0.011	0.000	0.026	0.047	0.024	0.138	0.087	0.023	0.007	0.033	0.031	0.020	0.021	0.169	0.059
mel	yak	0.044			0.020	0.009	0.292		0.000	0.116	0.098	0.060	0.108	0.171		0.104		0.249	0.241
mel	ere	0.049			0.010	0.006			0.220	0.168	0.225	0.060	0.118			0.159		0.249	0.205
mel	ana	0.110					0.143					0.229		0.243					
mel	pse	0.142														0.235			
mel	per	0.139														0.219			
mel	wil	0.208																	
mel	vir	0.237							0.392							0.350			
mel	moj	0.220							0.406					0.584					
mel	gri	0.246																	
sim	sec	0.005	0.021	0.008	0.082	0.000	0.035	0.045	0.018	0.043	0.005	0.011	0.005	0.016	0.015	0.015	0.012	0.164	0.010
sim	yak	0.038			0.064	0.009	0.302		0.004	0.085	0.117	0.056	0.107	0.194		0.105		0.221	0.198
sim	ere	0.045			0.054	0.006			0.224	0.166	0.217	0.058	0.116			0.111		0.221	0.178
sim	ana	0.101					0.273					0.229		0.212					
sim	pse	0.145														0.236			
sim	per	0.141														0.222			
sim	wil	0.203																	
sim	vir	0.231							0.354							0.320			
sim	moj	0.216							0.364					0.588					
sim	gri	0.239																	
sec	yak	0.035			0.019	0.009	0.282		0.019	0.090	0.072	0.063	0.108	0.138		0.092		0.094	0.208
sec	ere	0.045			0.019	0.006			0.220	0.157	0.211	0.061	0.121			0.108		0.094	0.174
sec	ana	0.100					0.303					0.244		0.214					
sec	pse	0.144														0.233			
sec	per	0.140														0.220			
sec	wil	0.201																	
sec	vir	0.231							0.388							0.319			
sec	moj	0.216							0.382					0.581					
sec	gri	0.239																	
yak	ere	0.051			0.025	0.003			0.230	0.146	0.193	0.049	0.039			0.076		0.000	0.043
yak	ana	0.100					0.283					0.231		0.307					
yak	pse	0.141														0.176			
yak	per	0.137														0.206			
yak	wil	0.204																	
yak	vir	0.237							0.347							0.310			
yak	moj	0.220							0.392					0.610					
yak	gri	0.242																	
ere	ana	0.102										0.217							
ere	pse	0.136														0.274			
ere	per	0.135														0.267			
ere	wil	0.204																	
ere	vir	0.231							0.403							0.420			
ere	moj	0.220							0.410										
ere	gri	0.235																	
ana	pse	0.159																	
ana	per	0.158																	
ana	wil	0.225																	
ana	vir	0.252																	
ana	moj	0.226												0.617					
ana	gri	0.252																	
pse	per	0.004														0.012			
pse	wil	0.174																	
pse	vir	0.224														0.300			
pse	moj	0.205																	
pse	gri	0.226																	
per	wil	0.173																	
per	vir	0.221														0.281			
per	moj	0.207																	
per	gri	0.227																	
wil	vir	0.238																	
wil	moj	0.220																	
wil	gri	0.231																	
vir	moj	0.144							0.264										
vir	gri	0.176																	
moj	gri	0.176																	

species 1	species 2	Adh	Invader5		Invader6		Max		Mdg1		Mdg3		Micropia		Ninja		Nobel		Nomad	
			I	LTR	I	LTR	I	LTR	I	LTR	I	LTR	I	LTR	I	LTR	I	LTR	I	LTR
mel	sim	0.019	0.558	0.071	0.012	0.011	0.004	0.003	0.027	0.007	0.017	0.003	0.005	0.006	0.023	0.009			0.007	0.014
mel	sec	0.017	0.535	0.065	0.012	0.008	0.009	0.016	0.014	0.034	0.019	0.003	0.005	0.014	0.017	0.009			0.007	0.017
mel	yak	0.044			0.006	0.018	0.014	0.020	0.032	0.091	0.009	0.004	0.071	0.126	0.021	0.012			0.117	
mel	ere	0.049	0.253	0.186					0.092	0.164			0.075	0.095	0.122	0.222			0.232	
mel	ana	0.110													0.120				0.338	
mel	pse	0.142													0.188					
mel	per	0.139													0.188					
mel	wil	0.208																		
mel	vir	0.237											0.376	0.503	0.230					
mel	moj	0.220			0.298								0.329							
mel	gri	0.246																		
sim	sec	0.005	0.160	0.019	0.008	0.013	0.005	0.003	0.021	0.019	0.012	0.003	0.004	0.000	0.000	0.012			0.007	0.008
sim	yak	0.038			0.006	0.019	0.006	0.047	0.041	0.094	0.017	0.060	0.076	0.129	0.019	0.012			0.101	
sim	ere	0.045	0.385	0.216					0.101	0.179			0.078	0.101	0.162	0.232			0.202	
sim	ana	0.101													0.134				0.414	
sim	pse	0.145													0.188					
sim	per	0.141													0.188					
sim	wil	0.203																		
sim	vir	0.231											0.353	0.509	0.274					
sim	moj	0.216			0.282								0.321							
sim	gri	0.239																		
sec	yak	0.035			0.005	0.019	0.012	0.043	0.088	0.084	0.016	0.059	0.078	0.121	0.018	0.012			0.063	
sec	ere	0.045	0.431	0.206					0.090	0.166			0.073	0.096	0.133	0.237			0.220	
sec	ana	0.100													0.091				0.312	
sec	pse	0.144													0.170					
sec	per	0.140													0.138					
sec	wil	0.201																		
sec	vir	0.231											0.378	0.503	0.171					
sec	moj	0.216			0.284								0.326							
sec	gri	0.239																		
yak	ere	0.051							0.090	0.175			0.076	0.084	0.123	0.239			0.280	
yak	ana	0.100													0.151		0.261		0.361	
yak	pse	0.141													0.188		0.251			
yak	per	0.137													0.149		0.203			
yak	wil	0.204																		
yak	vir	0.237											0.385	0.525	0.251					
yak	moj	0.220			0.267								0.343							
yak	gri	0.242																		
ere	ana	0.102													0.397				0.385	
ere	pse	0.136													0.407					
ere	per	0.135													0.198					
ere	wil	0.204																		
ere	vir	0.231											0.426	0.500	0.281					
ere	moj	0.220											0.413							
ere	gri	0.235																		
ana	pse	0.159													0.201		0.305			
ana	per	0.158													0.177		0.241			
ana	wil	0.225																		
ana	vir	0.252													0.274					
ana	moj	0.226																		
ana	gri	0.252																		
pse	per	0.004													0.010		0.086	0.020		
pse	wil	0.174																		
pse	vir	0.224													0.255					
pse	moj	0.205																		
pse	gri	0.226																		
per	wil	0.173																		
per	vir	0.221													0.269					
per	moj	0.207																		
per	gri	0.227																		
wil	vir	0.238																		
wil	moj	0.220																		
wil	gri	0.231																		
vir	moj	0.144											0.348							
vir	gri	0.176																		
moj	gri	0.176																		

species 1	species 2	Adh	Osvaldo	Quasimodo		Quasimodo2		Roo		RooA		Rover	Stalker2		Stalker4		Tabor		Tabor DA	
			I	I	LTR	I	LTR	I	LTR	I	LTR	I	I	LTR	I	LTR	I	LTR	I	LTR
mel	sim	0.019	0.122		0.136	0.005	0.009	0.005	0.002	0.049	0.031	0.080	0.001	0.000	0.018	0.031	0.002	0.006		
mel	sec	0.017	0.086		0.131	0.011	0.021	0.000	0.005	0.055	0.026	0.074	0.002	0.000	0.016	0.012	0.002	0.006		
mel	yak	0.044	0.282					0.007	0.002	0.060	0.061	0.056		0.118	0.048	0.195		0.028		
mel	ere	0.049	0.241	0.028	0.099			0.043	0.021	0.062	0.033	0.041		0.118	0.095	0.117				
mel	ana	0.110	0.308					0.307												
mel	pse	0.142	0.339					0.639							0.134	0.323				
mel	per	0.139	0.294					0.627							0.106	0.337				
mel	wil	0.208													0.149					
mel	vir	0.237													0.100	0.283				
mel	moj	0.220	0.430							0.317					0.225	0.243				
mel	gri	0.246	0.451																	
sim	sec	0.005	0.043		0.016	0.006	0.015	0.005	0.000	0.024	0.006	0.016	0.001	0.002	0.023	0.005	0.002	0.004		
sim	yak	0.038	0.201					0.012	0.005	0.042	0.052	0.120		0.126	0.056	0.192		0.022		
sim	ere	0.045	0.194		0.252			0.145	0.021	0.037	0.020	0.047		0.116	0.108	0.105				
sim	ana	0.101	0.217					0.333												
sim	pse	0.145	0.311					0.614							0.151	0.349				
sim	per	0.141	0.265					0.561							0.115	0.356				
sim	wil	0.203													0.154					
sim	vir	0.231													0.124	0.286				
sim	moj	0.216	0.389							0.297					0.244	0.215				
sim	gri	0.239	0.417																	
sec	yak	0.035	0.219					0.007	0.007	0.037	0.053	0.076		0.123	0.049	0.178		0.025		
sec	ere	0.045	0.209		0.230			0.021	0.024	0.042	0.029	0.039		0.113	0.103	0.106				
sec	ana	0.100	0.299					0.293												
sec	pse	0.144	0.330					0.640							0.152	0.322				
sec	per	0.140	0.280					0.628							0.105	0.344				
sec	wil	0.201													0.153					
sec	vir	0.231													0.108	0.251				
sec	moj	0.216	0.329							0.303					0.228	0.136				
sec	gri	0.239	0.305																	
yak	ere	0.051	0.226					0.031	0.019	0.026	0.023	0.006		0.037	0.115	0.139				
yak	ana	0.100	0.310					0.308												
yak	pse	0.141	0.338					0.640							0.166	0.374				
yak	per	0.137	0.315					0.623							0.131	0.425				
yak	wil	0.204													0.158					
yak	vir	0.237													0.127	0.341				
yak	moj	0.220	0.362							0.299					0.247	0.409				
yak	gri	0.242	0.386																	
ere	ana	0.102	0.319					0.331												
ere	pse	0.136	0.342					0.643							0.149	0.388				
ere	per	0.135	0.319					0.638							0.116	0.404				
ere	wil	0.204													0.142					
ere	vir	0.231													0.114	0.325				
ere	moj	0.220	0.400							0.302					0.246	0.313				
ere	gri	0.235	0.435																	
ana	pse	0.159	0.304					0.672												
ana	per	0.158	0.221					0.649												
ana	wil	0.225																	0.126	
ana	vir	0.252																		
ana	moj	0.226	0.314																	
ana	gri	0.252	0.171																	
pse	per	0.004	0.095					0.044							0.025	0.249				
pse	wil	0.174													0.119					
pse	vir	0.224													0.059	0.112				
pse	moj	0.205	0.338												0.215	0.466				
pse	gri	0.226	0.317																	
per	wil	0.173													0.123					
per	vir	0.221													0.052	0.246				
per	moj	0.207	0.307												0.233	0.493				
per	gri	0.227	0.298																	
wil	vir	0.238													0.112					
wil	moj	0.220													0.111					
wil	gri	0.231																		
vir	moj	0.144													0.227	0.347				
vir	gri	0.176																		
moj	gri	0.176	0.331																	

species 1	species 2	Adh	Tirant		Tram		Transpac		TV1		Zam	
			I	LTR	I	LTR	I	LTR	I	LTR	I	LTR
mel	sim	0.019	0.018	0.029			0.000	0.000			0.023	0.024
mel	sec	0.017	0.012	0.041			0.001	0.000			0.023	0.017
mel	yak	0.044	0.053				0.086	0.216			0.067	0.050
mel	ere	0.049	0.192				0.166				0.158	0.044
mel	ana	0.110					0.156					
mel	pse	0.142										
mel	per	0.139										
mel	wil	0.208										
mel	vir	0.237										
mel	moj	0.220									0.614	
mel	gri	0.246										
sim	sec	0.005	0.028	0.022			0.001	0.000			0.020	0.010
sim	yak	0.038	0.200				0.086	0.212			0.084	0.141
sim	ere	0.045	0.214				0.288				0.094	0.141
sim	ana	0.101					0.166					
sim	pse	0.145										
sim	per	0.141										
sim	wil	0.203										
sim	vir	0.231										
sim	moj	0.216									0.618	
sim	gri	0.239										
sec	yak	0.035	0.175				0.085	0.208			0.080	0.129
sec	ere	0.045	0.164				0.178				0.094	0.130
sec	ana	0.100					0.136					
sec	pse	0.144										
sec	per	0.140										
sec	wil	0.201										
sec	vir	0.231										
sec	moj	0.216									0.586	
sec	gri	0.239										
yak	ere	0.051	0.092				0.209				0.011	0.008
yak	ana	0.100					0.194					
yak	pse	0.141										
yak	per	0.137										
yak	wil	0.204										
yak	vir	0.237										
yak	moj	0.220									0.616	
yak	gri	0.242										
ere	ana	0.102					0.297					
ere	pse	0.136										
ere	per	0.135										
ere	wil	0.204										
ere	vir	0.231										
ere	moj	0.220									0.615	
ere	gri	0.235										
ana	pse	0.159										
ana	per	0.158										
ana	wil	0.225										
ana	vir	0.252										
ana	moj	0.226										
ana	gri	0.252										
pse	per	0.004			0.146	0.003						
pse	wil	0.174										
pse	vir	0.224										
pse	moj	0.205										
pse	gri	0.226										
per	wil	0.173										
per	vir	0.221										
per	moj	0.207										
per	gri	0.227										
wil	vir	0.238										
wil	moj	0.220										
wil	gri	0.231										
vir	moj	0.144							0.009	0.005		
vir	gri	0.176										
moj	gri	0.176										