EDINBURGH
UNIVERSITY
LIBRARY

Shelf Mark DARWIN LIBRARY

WANG PhD. 2010

University of Edinburgh
30150 025252393

Downie Allison Downie Ltd
Unit H • Purdon Street
Glasgow • G11 6AF
0141 339 0333
www.dadbookbinders.com

Alignment and analysis of noncoding DNA sequences in *Drosophila*

Jun Wang

Thesis presented for the degree of Doctor of Philosophy

University of Edinburgh

2008

EX BIBL UNIV EDINBURGH

Table of contents

Acknowledgements

I thank Toby Johnson for generously offering theoretical and technique help for the project of MCALIGN2. Without his guidance and help, the MCALIGN2 programming would be a long, slow process. I also want to give my genuine and massive thanks to Daniel Halligan. For the studies of transposable elements, he has given me numerous help in relation to the research direction and data analysis ideas. He has also helped me with the English correction of published papers and thesis writing. With his help I could carry out the study of molecular evolution of *Drosophila* smoothly and firmly. I also want to thank my supervisor Prof. Peter Keightley for his supervision and guidance of my PhD projects. He also offered great help in paper and thesis writing. I also thank Daniel Gaffney and Brian Charlesworth for reviewing my study on INE-1 elements.

Throughout my PhD study, there are many mates and friends I want to thank for the support they have given me, particularly when I was down and low. I thank Dominic, Xiaolu, Rachel, David, Telina, Priscilla, Feifei, Jessie, Isabelle and my officemates, Silvia, Dario. Without them and their support, my work would be much less exciting, and I would not have much enthusiasm to carry out my work. Thanks again to them. You are aces.

Finally I want to thank my dear parents. Thanks to them for raising and educating me to be responsible and motivated. Thanks to them for their unselfish support on my life, education and everything. This thesis is attributed to them.

Publications

- Wang, J., Keightley, P. D. and Johnson, T. (2006). MCALIGN2: faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. *BMC Bioinformatics* **7**: 292.

- Wang, J., Keightley, P. D. and Halligan, D. L. (2007). Effect of divergence time and recombination rate on molecular evolution of Drosophila INE-1 transposable elements and other candidates for neutrally evolving sites. *J Mol Evol* **65**: 627-639.

- Wang, J., Keightley, P.D. and Halligan, D. L. (2008). Transposable elements are more than just genomic parasites: a case study in *Drosophila*. In preparation.

Abstract

In many higher eukaryotic organisms, a substantial portion of the genome is comprised of noncoding DNA, whose potential contribution to gene functions and genome evolution is still relatively unknown. Since the completion and availability of genome sequences of numerous species, comparative genomic analysis has become a powerful tool to reveal the functional significance of noncoding sequences, by searching for signs of sequence conservation. It has been shown that a large fraction of noncoding sequences are evolutionary constrained in *Drosophila*, and transposable elements (TEs) are one of the major components of noncoding DNA. To compare and analyze genomic sequences within and among species, accurate alignment of orthologous sequences from related species is a vital first step.

In this thesis, a fast accurate global pairwise alignment of noncoding DNA sequences, MCALIGN2, is developed based on explicit models of indel evolution. A pair-hidden-Markov-Model (pair-HMM) of seven states and a Golden-Section-Search algorithm are employed in this method to search for the most probable alignment between two homologous sequences. MCALIGN2 outperforms other available alignment methods in simulations for all combinations of parameter values considered. This method is then used to align and analyze noncoding DNA sequences in *Drosophila*.

Comparative genomic analysis in this thesis shows that INE-1 elements, one of the most abundant TEs in *Drosophila*, along with sites within short introns and fourfold degenerate sites are the fastest evolving nucleotides in the genomes of *Drosophila melanogaster*, *D. simulans* and *D. sechellia*. Fourfold sites tend to be evolving

(relatively) slightly more slowly than the other two classes of nucleotides, probably due to selection acting on protein translation efficiency. The observed substitution rate in these fastest evolving sites appears to be strongly influenced by the recombinational environment in which they are located. This rate may be influenced by several factors including ancestral polymorphisms, variation in mutation rate, natural selection and random genetic drift. The relative importance of these factors varies depending on the time since speciation.

This thesis also fully investigates the distribution and rates of evolution of three major TE classes (LTR, non-LTR retrotransposons and DNA transposons) in the *Drosophila* euchromatic genome using a gene-centric approach. The study demonstrates that LTR elements outnumber non-LTR and DNA elements in all intergenic, intronic and exonic regions, and LTR elements also show relatively lower mean divergences than the other two classes between *D. melanogaster* and *D. yakuba*. The *roo_I* elements, a *Pao* family retrotransposon, appear to be evolving the most slowly among LTR elements. The study also shows that there are fewer TE insertions in *Drosophila* promoter regions than in intergenic regions that are not close to coding sequence boundaries. TEs in promoter regions were also found to have lower interspecies mean divergences than those in distal regions. These findings suggest that some TEs, rather than being "junk" and "selfish", may be conserved between species, and therefore, play vital roles in gene regulation and host genome evolution.

Chapter 1.

Introduction

1.1. DNA sequence evolution

The emergence of automated DNA sequencing has triggered a massive growth in the volume of sequence data deposited in public databases. This allows us to analyse DNA sequence data on a very large scale, both within and between species. Meanwhile, greater computational power also helps process a huge amount of data in a substantially shorter time, compared to a decade ago. DNA sequences are generally comprised of protein-coding DNA sequences and non-protein-coding DNA sequences. Protein-coding DNA sequences are DNA sequences that are transcribed into mRNA and subsequently translated into proteins (amino acid sequences). Translation starts at the translation initiation site and proceeds to a stop signal. Every triplet of adjacent nucleotides (i.e., a codon) in a protein-coding sequence specifies a single amino acid. Non-protein-coding DNA sequences, here referred to as just noncoding DNA, describe DNA which does not contain instructions for making proteins. Noncoding DNA sequences include introns that are located within coding regions, and intergenic sequences between genes.

1.1.1. Mutation and recombination

Mutations are changes to the nucleotide sequence of the genetic material of an organism. Mutations can be caused by copying errors in the genetic material or incorrect DNA repair during cell division or by external stimuli, such as exposure to ultraviolet or

ionizing radiation, chemical mutagens, or viruses. Mutations can also occur under cellular control during processes such as hypermutation without any impact of external factors (Griffiths et al. 1999, Chapter 7). In multicellular organisms, mutations can be subdivided into germ line mutations (which can be passed on to descendants) and somatic mutations (which are not transmitted to descendants). The former appear to be relatively more interesting from the aspect of sequence evolution, since they provide the source of heritable variation. Mutations can also be classified into (1) base-pair substitutions, the replacement of one nucleotide by another; (2) deletions, the removal of one or more nucleotides from the DNA; (3) insertions, the addition of one or more nucleotides into the DNA; (4) inversions, the chromosome rearrangement in which a segment of a chromosome is reversed end to end (Li 1997, p.23-29).

Base-pair substitutions can be classified into transitions and transversions. Transitions are substitutions between A and G (purines) or between C and T (pyrimidines). Transversions are substitutions between a purine and a pyrimidine (e.g., A→T and A→C). Thus, there are four types of transitions and eight types of transversions. In protein-coding regions, a substitution is said to be synonymous or silent if it causes no amino acid change, otherwise, it is nonsynonymous. Deletions and insertions are collectively referred to as indels (or gaps) in the sequence alignment, because one cannot tell whether a deletion occurred in one descendant sequence or an insertion occurred in the other descendant sequence in a pairwise sequence alignment without information from the ancestral sequence.

In addition to mutations, recombination can also cause changes to the genetic material. It refers to the process by which two chromosomes pair up and exchange their DNA content. Recombination includes chromosomal crossover and gene conversion. In eukaryotes, chromosomal crossover (or crossing over) generally occurs during the formation of gametes or meiosis. It involves two chromosomes of a homologous pair exchanging sections of their DNA reciprocally. If one is considering a metaphase in meiosis where each chromosome consists of two (initially identical) sister chromatids, crossing over can then occur between one chromatid of one of the homologous chromosomes and one chromatid of the other homologous chromosome. For example, a pair of homologous chromosomes consists of two linked diploid loci. One chromosome has two sister chromatids *AB*, and the other chromosome consists of two other sister chromatids (different from the first chromosome) *ab*. Crossing over between one chromatid *AB* of the first chromosome and one chromatid *ab* of the second chromosome will yield a 1:1:1:1 ratio in the four haploid products of this meiosis, i.e., 1*AB*:1*Ab*:1*aB*:1*ab*, instead of 2*AB*:2*ab*. Gene conversion refers to the process by which one section/locus of the genetic material is transferred from one of the paired chromosomes to the other, but the donating chromosome stays unchanged. Thus, this is a non-reciprocal change (e.g., the B→b change at a locus, but no b→B change). Gene conversion also occurs during meiotic division, and, as mentioned, it generally involves the breakage of one of the paired chromosomes and the invasion of the genetic content from the other chromosome to the homologous DNA strand of the broken chromosome. In the case of gene conversion, the segregation at the *B* locus will be 3*B*:1*b* instead of 2*B*:2*b*, while segregation at the A locus will still be 2*A*:2*a*. Gene conversion and

crossing-over are often associated events. For example, the correlation between these two events could be 50% or higher in yeast (Borts and Haber 1987; Petes et al. 1991).

Therefore, recombination also creates genetic variation as mutations do. It does so by rearranging genetic content based on the existing genetic material, while mutations create genetic variation by introducing new genetic material to the organism.

1.1.2. Natural selection and genetic drift

Mutations and recombination create variations in the gene pool of a population, and their fates then will be determined by natural selection and/or random genetic drift. There has been heated debate between selectionists and neutralists since the neutral theory first came out (Kimura 1968). Their relative importance of selection and random drift depends on the effects of new mutations (or recombination) on host fitness.

The selectionist hypothesis proposes that most new mutations in a population will reduce the host fitness, and are deleterious. Such mutations will be selected against and will be eventually eliminated from the population by negative or purifying selection. However, it is possible that a new mutation favours its host by introducing a selective advantage over other individuals. Such a mutation may eventually become fixed in the population through positive selection. It is also possible that mutant type alleles and the wild type alleles co-exist at a locus in the population with different frequencies (i.e., polymorphism), resulting in selective advantages for the population. This is maintained by balancing selection. Natural selection acts on the phenotype, or the observable characteristics of an organism, and it leads to changes in allele frequencies over time.

The neutralist hypothesis proposes that the vast majority of the changes that take place in evolution are selectively neutral or nearly so (i.e., slightly deleterious) (Kimura 1968; Ohta 2002). Their fates will be largely determined by random genetic drift. Random genetic drift refers to the process of randomly sampling gametes during reproduction of a population. If the population has $2N$ adults, and if the effective population size $N_e = N$, then the population will create a vast number of gametes, of which $2N$ are sampled to be found in the adults of the next generation. The assumption is made here that each individual in the population is equal in fitness, and selection and other evolutionary forces are absent. Thus, the chance that each allele can be passed down to the next generation solely depends on stochastic sampling. During the process of random drift, some alleles are lost in each generation by chance, and eventually, each allele at a locus will become a copy of just one of the ancestral alleles in generation 0. The probability that a particular allele in generation 0 will become fixed in the population is simply the fraction of that allele in the population initially, $\frac{1}{2N}$. Considering a simple case of one locus with two alleles, A_1 and A_2, with frequencies p and $q = 1 - p$, respectively, the probability that A_1 will become fixed in the population is equal to the initial allele frequency of A_1, simply just p. According to the neutral hypothesis, polymorphism is a temporary state that is still undergoing random genetic drift. Polymorphism is more frequent for large populations than for small ones, since random genetic drift tends to be a slower process in larger populations.

Selection will still operate on a new mutation that is nearly neutral, or slightly deleterious, but chance effects are a stronger determinant of eventual fixation than is selection. The relative importance of natural selection and genetic drift in a population varies depending on the strength of the selection and the effective population size N_e, defined as the size of an idealized population that would have the same effect of random sampling on gene frequency as that in the actual population (Wright 1931).

The neutral theory is widely used as a "null model" of sequence evolution. The neutrality test compares the actual/observed number of differences (substitutions) between two sequences and the expected number that neutral theory predicts given the independently estimated divergence time (e.g., the expected number of differences for putatively neutrally evolving sites). If the actual number of differences is (much) less than the predicted, the null hypothesis is disproved. We then may reasonably assume that selection has acted upon the sequences in question. This is the basis for identifying evolutionary constraint (Keightley and Gaffney 2003; Halligan et al. 2004; Andolfatto 2005; Halligan and Keightley 2006).

1.1.3. The role and survival of noncoding DNA

In eukaryotes, genome sizes (termed the C-value, denoting the total amount of DNA in a haploid genome) vary dramatically among species, from ~2.5Mb in *Thalassiosira pseudonana* to ~670,000Mb in *Amoeba dubia* (Sparrow et al. 1972; Cavalier-Smith 1985; Li and Graur 1991; Armbrust et al. 2004). However, there is no clearly established relationship between the interspecific variation in genome sizes and organismic

complexity or the likely number of genes encoded by the organism (Li 1997, p.381; Gregory 2005, p.7-10). For example, several unicellular protozoans (e.g., amoebae) possess much more DNA than mammals (up to ~200-fold) (Li and Graur 1991). Moreover, organisms that seem similar in genetic complexity possess vastly different C-values. For example, the C-values of flies and locusts are ~180Mb and ~9,300Mb (Li and Graur 1991), respectively, although they are thought to have similar amounts of "genetic information". This lack of correlation between C-values and the presumed amount of genetic information within genomes is known as the C-value paradox or the C-value enigma (Thomas 1971; Gregory 2001). It has become clear that the C-value enigma appears to be mostly due to the presence of noncoding DNA and its variability in quantity among organisms (Gregory 2005, p.9-10 and p.27; Li 1997, p.381-384).

Large proportions of noncoding DNA are made up of transposable elements in most eukaryotic genomes. Transposable elements (TEs), also known as "jumping genes", could play an important role in accounting for the C-value enigma. One should note that much TE DNA is coding, although coding for the proteins required for TE movement. The proportion of the genome taken up by TEs also seems to vary widely among taxa. For instance, TEs comprise ~50% and ~40% of human and mouse DNA, respectively (International human genome sequencing consortium 2001; Waterston et al. 2002); at least 50% of the maize genome (SanMiguel et al. 1996) and approximately 90% of the genome of some species of lilies (Flavell 1986). In species with relatively smaller genomes, the percentage of TEs in the genome is relatively less, e.g., 6% in *C. elegans* (Waterston and Sulston 1995) and less than 10% in *D. melanogaster* (Kaminka et al.

2002). Generally, it is assumed that TEs are relatively more important determinants of genome size (C-value) in large genomes compared to small genomes (Kidwell 2002).

For a long time, scientists believed that noncoding DNA was useless junk (Ohno 1972), or "selfish DNA" replicating more efficiently than coding DNA (Orgel and Crick 1980). However, there would be some drawbacks for the genome to maintain a large amount of "junk" or "selfish" DNA. A large amount of noncoding DNA tends to exhibit greater sensitivity to evolutionary changes, although most of the changes may be selectively neutral or do not become fixed in the population. Furthermore, maintaining and replicating a large amount of DNA may impose a burden on the organism (Li 1997, p.399; Lynch 2007, p32-35). Thus, it is possible that some noncoding DNA is maintained due to its evolutionary significance.

Indeed, various important roles of noncoding DNA have been discovered in recent years. First, there is evidence that some noncoding DNA (e.g., telomeres) is necessary for maintaining chromosomal structure and function (Sandell and Zakian 1994; Barinaga 1997; Harrington et al. 1997). Second, it has been shown that eukaryotic noncoding DNA are crutial to maintain the secondary DNA structure (Cavalier-Smith 1985; Beaton and Cavalier-Smith 1999). Third, noncoding DNAs may play a vital role in the regulation of gene expression during development (Ting 1995; Vandendries et al. 1996; Keplinger et al. 1996; Kohler et al. 1996). This is supported by a great body of evidence demonstrating the role of noncoding DNA as promotors/enhancers (or silencers) for transcription (or suppression of transcription) of proximal genes since several decades ago. Recent discovery includes the evidence shown in Nikolajczyk et al. (1996),

Tanimoto et al. (1996), Tiffany et al. (1996), Bouhassira et al. (1997), Handen and Rosenberg (1997) and van de Lagamaat et al. (2003). Fourth, some of the noncoding DNA establishes the correct reading frame for translation (Trifonov 1989). Fifth, in recent years, there has been growing evidence that noncoding RNA (ncRNA) genes (that are transcribed into noncoding RNA) are far more widespread than was expected before, e.g., in mammals (Eddy 2002; Hüttenhofer et al. 2002), and they may be very important in regulating gene expression (acting as molecular switches) (Eddy 2002; Mattick 2004; Hüttenhofer et al. 2005). In addition, 3' and 5' untranslated regions (UTR) are crucial for regulating gene expression. The 3' and 5' UTRs provide the binding sites for certain trans-acting binding proteins to regulate gene expression. Well characterized examples of this include the lipoprotein lipase gene (Ranganathan et al. 1997), the glucose transporter gene (McGowan et al. 1997) and the glutathione peroxidase and phospholipid-hydroperoxide glutathione peroxidase genes (Bermano et al. 1996).

More recently, since the completion and availability of genome-scale sequence for numerous species, comparative genomic analysis has revealed many highly conserved nongenic sequences (CNGs) whose functional significance is still poorly understood, particularly in mammals (Frazer et al. 2001; Dermitzakis et al. 2002, 2003, 2004; Bejerano et al. 2004). These CNGs are subject to strong negative selection, and are highly selectively constrained between related species (Drake et al. 2005; Keightley et al. 2005). Noncoding DNA sequences, contrary to the statements of their being "junk" or "selfish", are thus not useless, but at least some proportions are, in fact, required for genomic functionality.

1.1.4. Principle of sequence comparison

Comparison is a powerful tool in biological science. As mentioned above, recent comparative genomic-based strategies have begun to help identify functional sequences based on their high levels of evolutionary conservation, resting on the hypothesis that important/functional biological sequences are conserved between species due to evolutionary constraints (Nobrega and Pennacchio 2003). In other words, functionally important parts of the genome are expected to evolve more slowly than those lacking function if newly arising mutations are deleterious to gene function (Shabalina and Kondrashov 1999; Andolfatto 2005). Technological progress in DNA sequencing has resulted in the generation of a large dataset of genomic sequence information for numerous species. This has allowed us to produce genome-wide alignments to compare and contrast the evolution and content of genomes of related species. Such comparisons are vital to identify blocks of conserved sequences over evolutionary time, and such evolutionary conservation has been a powerful guide in revealing functional elements of noncoding DNA (Hardison 2000; Pennacchio and Rubin 2001; Nobrega and Pennacchio 2003). Sequence alignment is the first step in genomic comparison.

1.2. Sequence alignment

Sequence alignment is one of the most important issues in comparative genomic analysis. It is the process of lining up homologous bases, which also involves identifying the locations for insertions and deletions (indels). Sequence alignment relies on the

assumption that the two sequences under study have been derived from a common ancestral sequence, and have been evolving independently ever since.

1.2.1. Alignment categories

There are several categories of alignment: pairwise alignment is the comparison of two sequences, while multiple alignment is a natural extension of pairwise alignment to incorporate more than two sequences at a time. Sequence alignment can also be performed globally or locally. Global alignment assumes that the two sequences are similar over their entire length, and attempts to match them to each other from end to end based on global similarities derived from the ancestral sequence. Local alignment is an alignment algorithm to search for regions of local similarity between sequences. The most common situation involves a query sequence x that is much shorter than a second sequence y, treated as a target database sequence (e.g., the entire genome sequence). Local alignment will look for the best alignment between sequence x and subsequences of y. It is usually a more sensitive way to detect similarity than global alignment, especially when comparing two very highly diverged sequences (Durbin et al. 1998, p.22). However, one should not simply compare local alignment with global alignment from the aspect of alignment accuracy, since they were designed for different alignment tasks. This thesis mainly focuses on the introduction to and implementation of global pairwise alignment.

1.2.2. Pairwise alignment

For a given pairwise alignment, there are three types of aligned pairs: (1) matched base

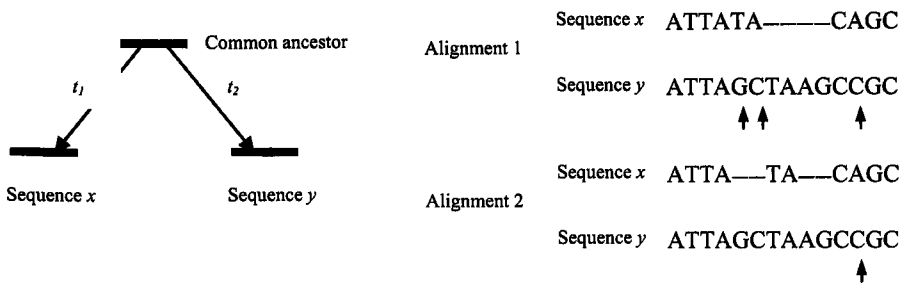


Figure 1.1. – Sequence x and sequence y have diverged from the common ancestor for divergence time t_1 and t_2, respectively. For any two sequences, we can come up with many alignments when gaps are included. There are 3 substitutions and 1 gap in alignment 1, and 1 substitution and 2 gaps in alignment 2. Substitutions (mismatches) are indicated by arrows.

pairs, (2) mismatched base pairs, and (3) pairs consisting of a base from one sequence and a gap from the other sequence, denoted by "-" (Figure 1.1). These three types of paired events correspond to the three basic evolutionary processes for sequence evolution. Matched base pairs suggest that no point substitutions have occurred at the target site since the divergence between the two sequences. Mismatched base pairs suggest that (at least) one substitution has occurred at the target site in either sequence. A gap state implies that an indel has occurred in one of the two sequences.

One should note that when we include gaps in the alignment, there could be many possible alignments for any two sequences (e.g., alignment 1 and alignment 2, Figure 1.1). We are faced with a choice between having more point substitutions and having more gap events. We must then find a scoring system with which to compare gaps and substitutions. Such a scoring system will include scores for matches/mismatches, and scores for gaps (i.e., gap penalty).

1.2.3. Available alignment methods

Two famous alignment algorithms have been developed decades ago, to deal with global alignment and local alignment, respectively. They are termed the Needleman-Wunsch algorithm (1970) and the Smith-Waterman algorithm (1981). They have become the basis for the following new algorithms.

The Needleman-Wunsch approach

Needleman and Wunsch (1970) suggested a progressive building of an alignment, based on dynamic programming. This algorithm is guaranteed to find the optimal alignment for two similar-sized sequences. The dynamic programming procedure can be carried out by first constructing a score-based matrix F between the two sequences of length x and y, respectively. Matrix F will be a $(x+1) \times (y+1)$ matrix, as the score at $F(0, 0)$ is initialized as 0. Three basic events in a pairwise alignment, matches/mismatches/indels, are each given a score or a score scheme. The procedure starts from the beginning of each sequence $F(0, 0)$, and then fills the matrix by moving to positions next to the previous position from left to right and from top to bottom. The score at the target position is based on scores at previous positions plus the transition scores between them. The maximum score of them is then chosen to be stored at the target position. For example, at the target position $F(m, n)$, the score there is partially determined by scores at the three previous positions $F(m-1, n-1)$, $F(m-1, n)$ and $F(m, n-1)$. The transition from $F(m-1, n-1)$ to $F(m, n)$ is to enter a match state if the mth base of one sequence is the same as the nth base of the other sequence, or a mismatch if they are

different bases. The transition from $F(m\text{-}1, n)$ to $F(m, n)$ is to enter a gap state of the nth base of one sequence aligning to a gap "-" of the other sequence, likewise for the transition from $F(m, n\text{-}1)$ to $F(m, n)$, entering the state of the mth base of the other sequence aligning to a gap. Thus, the score at position $F(m, n)$ is,

$$F(m,n) = \max \begin{cases} F(m-1,n-1) + s(m,n) \\ F(m-1,n) + d \\ F(m,n-1) + d \end{cases} \tag{1.1}$$

where $s(m, n)$ is the score for a match or a mismatch, and d is the score for a gap. Once the maximum score is placed, a pointer back to the position from which the maximum score was derived is also kept. Therefore, the memory to build the matrix F doubles. The building procedure continues filling the rest of the matrix recursively until the last position $F(x, y)$ is filled. The second part of the dynamic programming procedure is to start from position $F(x, y)$ and to trace back through the path that generated the maximum scores for each traceback position. At the end, the traceback procedure reaches the start of the matrix. The two sequences can then be aligned according to this path. Although the dynamic programming procedure provides a clear instruction of searching for the optimal alignment between two sequences, it does not provide any guidance of choosing parameters for a match/mismatch/indel.

In the Needleman-Wunsch approach, every match is given a score of 1, every mismatch a score of 0 and individual gaps a penalty score (Needleman and Wunsch 1970; Mount 2004, p.12). The most commonly used gap penalty models are the linear gap model (Equation 1.2) and affine gap penalty model (Equation 1.3),

$$w_k = bk, \tag{1.2}$$

$$w_k = a + b(k-1) \tag{1.3}$$

where w_k is the penalty for a gap of k bases, b is the gap-extension penalty and a is the gap-open penalty. The Needleman-Wunsch similarity score S between two aligned sequences is then (also shown in Mount 2004, p.12),

$$S = x - \sum w_k g_k \tag{1.4}$$

where x is the number of matched pairs and g_k is the number of gaps of $k$bp in length. Under the linear/affine gap penalty weight, the choice of a and b is a difficult problem, because the optimal alignment can be different for the two sequences, e.g., in Figure 1.1, given different sets of a and b. For example, if $a = 0.5$ and $b = 2$, the similarity score for alignment 1 is $S = 0.5$, and for alignment 2 is $S = 4$. Thus, alignment 2 is favoured over alignment 1. However, if $a = 3$ and $b = 0.5$, then $S = 2.5$ for alignment 1, whereas $S = 2$ for alignment 2. Alignment 1 becomes favoured over alignment 2. A quality global alignment is vital to infer the actual differences (substitutions) between the two homologous sequences.

The Smith-Waterman approach

An alignment is usually comprised of well-aligned subregions and relatively poorly-aligned subregions. In sequence comparison, well-aligned regions are usually the most biologically significant regions in DNA or protein sequences, while relatively poorly-aligned regions are made up of sequences that are less related compared to well-aligned

regions. Therefore, to locate these well-aligned subregions in a target database sequence is crucial for homology study. Smith and Waterman (1981) developed an important modification to the Needleman-Wunsch algorithm, now known as the Smith-Waterman algorithm, to look for high scoring local alignments. This approach finds the similarity between subsequences of a target database sequence and a relatively shorter query sequence. There could be many subsequences exhibiting similarity to the query sequence if the database sequence is long (e.g., the whole genome sequence), but their similarity scores would be different. Note that local alignment is not the focus of this project.

Heuristic methods

The Needleman-Wunsch and Smith-Waterman algorithms are guaranteed to find the optimal score according to the specified scoring scheme. However, they are not the fastest available sequence alignment methods, and to search with many different sequences and/or long sequences, time consumption becomes an important issue. Heuristic methods were introduced for this reason. However, some sensitivity will be sacrificed by heuristic approaches, and the optimal alignment could not be guaranteed, in particular when sequences are not closely related.

(1) CLUSTALW

The most widely used program for global sequence alignment is CLUSTALW (Thompson et al. 1994). CLUSTALW is a progressive multiple alignment tool based on a guide tree. It improves the sensitivity of progressive multiple sequence alignment

through some carefully designed strategies. For example, it uses different amino acid/nucleotide base substitution matrices at different alignment stages according to the divergence of the sequences to be aligned. It also dynamically chooses gap penalties in a position- and residue-specific manner (Thompson et al. 1994). For positions where gaps have been opened in early stages, gap penalties are reduced to encourage new sequences to have gaps in such positions. For regions where gaps appear to be relatively more frequent than others, residue-specific penalties are used to encourage gaps (Thompson et al. 1994). These strategies have been proved to be efficient to align protein sequences and simulated noncoding DNA sequences with conserved blocks (Thompson et al 1994; Pollard et al. 2004). However, these arbitrary strategies appears to perform poorly in aligning distantly related sequences, e.g., many noncoding DNA sequences (Keightley and Johnson 2004; Wang et al. 2006)

(2) BLAST and FASTA

The BLAST (Altschul et al. 1990; Altschul et al. 1997) and FASTA (Pearson and Lipman 1988) packages are two of the most widely used programs for local alignments (based on the Smith and Waterman algorithm), in the scenario of finding local high scoring alignments between a query sequence x and a relatively longer target sequence y, e.g., the whole genome sequence. The common steps of these two methods involve initially looking for short stretches of matched words (sequences), subsequently extending to search for good longer matched words and finally identifying gapped alignments (Durbin et al. 1998, p.33-34). A series of local similarity scores are calculated and ranked corresponding to each identified local alignment. The main

difference between the two methods is that FASTA realigns the highest scoring candidate matches using full dynamic programming (Pearson and Lipman 1988). This will give the FASTA package more sensitivity, but relatively slower speed, compared to BLAST (Durbin et al. 34).

(3) RepeatMasker

RepeatMasker is a widely used package to screen DNA sequences for interspersed repeats and low complexity DNA sequences (http://www.repeatmasker.org). It incorporates a sequence comparison program "cross_match", which was developed based on the Smith-Waterman-Gotoh algorithm (Smith and Waterman 1981; Gotoh 1982), to mask repeats in target sequences. The query sequences (i.e., repetitive sequences) are stored in the RepeatMasker library, and the newly discovered repeats or updated versions of canonical sequences are available at Repbase (http://www.girinst.org/). The program outputs annotations of all recognized interspersed or simple repeats in the masked database sequence(s), and optionally local alignments of the query with the matching repeats within the masked sequence(s). The program also provides options to restrict or ease the masking criteria.

(4) AVID

AVID (Bray et al. 2003 and 2004) is a global pairwise alignment program. The program firstly concatenates the two sequences (to be aligned) with the character *N* between them, and finds all maximal repeated substrings (matches) between the two sequences using a suffix tree. The program then starts an anchor selection process based

on the entire maximal match set. Short matches are removed and the rest of the set are selected as anchors using a variant of the Smith-Waterman algorithm. The anchors are then used to split long sequences into short sequences. Short sequences are then aligned by the Needleman-Wunsch algorithm using standard parameters. However, if there are still significant matches within any of the short sequences, the anchor selection process will be repeated within these sequences and these short sequences are split into even shorter ones that will be aligned by dynamic programming. At the end, the program orders and orients all aligned short sequences to form a whole global pairwise alignment between the two input sequences (Bray et al. 2003). This approach is fast and memory efficient, and is practical for sequence alignments of large genomic regions up to megabases long (Bray et al. 2003).

1.2.4. Statistical alignment

All of the methods discussed above are score-based methods, including a match/mismatch score for nucleotides or amino acids, and a cost function for indels. This means that parameters used to align the two sequences are predefined in advance of the alignment. This could cause some problems for biological sequence alignments. First, as discussed previously, the best alignment can vary, and will be strongly influenced by how we choose the parameters. Second, since score-based methods lack a proper statistical explanation for all parameters used in the alignment, it is difficult to assess the alignment reliability. Thus, there will be no statistical/probabilistic basis to compare several different alignments, let alone a biological basis.

The relevance of statistical approaches to evolutionary inference has long been recognized. Time-continuous Markov models for substitution processes were introduced more than three decades ago (i.e., the Jukes-Cantor model, Jukes and Cantor 1969). Compared to score-based approaches, the statistical (probabilistic) treatment to evolutionary processes (insertions/deletions/substitutions) can produce probabilistically and biologically meaningful parameters based on explicit models of evolution. These parameters are usually estimated by maximum likelihood (ML) or Bayesian techniques, in which all uncertainty of known and unknown parameters is accounted for. The Bayesian approach allows the assessment of the reliability of the alignment estimates by calculating the posterior probability. Furthermore, this also allows the comparison of different models of evolution and hypothesis testing (Lunter et al. 2004 and 2005).

The first evolutionary model for pairwise sequence alignment, termed the TKF91 model, was introduced by Thorne, Kishino and Felsenstein (1991). They proposed a time-reversible Markov model for insertions and deletions, as well as for point substitutions. This method uses a ML algorithm to estimate the evolutionary distance between two sequences and to obtain the most probable alignment. The TKF91 model was then modified to model indel events of more than one residue in length, roughly according to a geometric length distribution for each single indel event, becoming the TKF92 model (Thorne et al. 1992). In general, the advantages of a statistical approach to sequence alignment consist of the possibility of parameter inference, assessing uncertainty, hypothesis testing and model comparison (Chatfield 1995; Lunter et al. 2004), which are absent in score-based methods. However, challenges still remain for

the statistical approaches, e.g., to include more biological realism, and to incorporate variable substitution rates and more accurate indel evolutionary models (Lunter et al. 2004).

1.2.5. Non-coding sequence alignment

Accurate inferences of the function of noncoding DNA from comparative methods depend critically on correct alignments of noncoding sequences. However, the alignment of noncoding sequences is more difficult than aligning protein-coding sequences. There is usually little difficulty in producing convincing alignments of protein-coding sequences, because indels are relatively uncommon, usually occurring in multiples of three base pairs. Furthermore, protein-coding sequence alignment is also supported by the alignment of amino acid sequences that they encode. For noncoding DNA alignment, unless sequence divergence is low, indels can cause severe problems, by introducing too many gaps but too few nucleotide differences, or too few gaps but too many differences into the alignment (Figure 1.1). We must model indel events in noncoding DNA with more care, because the length distribution of indels can be quite different not only between species, but also between genomic regions. The TKF model, which assumes single-base pair indels only, appears to be inaccurate for noncoding DNA alignment, and heuristic methods that simply assign penalty scores to indels do not seem to deal with complex indel events properly, especially when divergence is high.

Keightley and Johnson (2004) developed a new approach MCALIGN for aligning noncoding DNA. The alignment is estimated assuming a model that allows an arbitrary

distribution of indel lengths. The distribution of indel lengths is first derived empirically from data in which unambiguous alignments have been made between closely related species (e.g., *Drosophila simulans* and *Drosophila sechellia*), as well as indel rate relative to nucleotide substitutions. This distribution can then be used to estimate alignments of sequences from relatively more distantly related species within the same group (e.g., *Drosophila melanogaster* and *Drosophila yakuba*). They then used a stochastic hill-climbing algorithm that searches for more probable alignments. They have compared their method with several widely-used heuristic approaches to global pairwise alignment, CLUSTALW (Thompson et al. 1994), AVID (Bray et al. 2003), DIALIGN (Morgenstern 1999), LAGAN (Brudno et al. 2003) and HANDEL (Holmes and Bruno 2001), using simulations. MCALIGN appeared to outperform other available methods (Keightley and Johnson 2004). This approach to modeling the distribution of indel lengths will be applied to a new method that this project has developed, but the new method will use a different optimizer.

1.3. The *Drosophila* genome project and searching for functional noncoding DNA

A large portion of the *D. melanogaster* genome was first sequenced by a consortium lead by Celera Genomics using the whole-genome shotgun sequencing method (Adam et al. 2000). This initial draft sequence, which contained many gaps and regions of low sequence quality, was later refined in the third release (Celniker et al. 2002). This assembly still had 44 gaps remaining. The 118.4Mb Release 4 euchromatic sequence

contained 6 chromosome arms (2L, 2R, 3L, 3R, 4 and X) with a total of 23 sequence gaps. The assembly was validated in collaboration with the Genome Science Centre at the British Columbia Cancer Agency in Vancouver, Canada, using fingerprint analysis of a tiling path of BACs spanning the genome (www.flybase.org). The ~120Mb euchromatic portion was thought to include the vast majority of the protein-coding genes. This result was built upon by a sequencing project for the euchromatic portion of the sister species *D. pseudoobscura* (Release 2), using a comparative sequence approach (Richards et al. 2005).

The genome sequence of *D. melanogaster* provided one of the first full genome models to study gene functions, genome structure and evolution (Misra et al. 2002; Celniker and Rubin 2003; Ashburner and Bergman 2005). 10 additional *Drosophila* genomes have also recently been sequenced and assembled (Myers et al. 2000; *Drosophila* 12 Genomes Consortium 2007). These species were chosen to span a wide variety of evolutionary distances, and the evolutionary divergence spanned among the 12 *Drosophila* species exceeds that of the entire mammalian radiation when generation time is taken into account (Stark et al. 2007). The wealth of choice of species, as well as genomic sequence volume has allowed comparative genomic analysis to be conducted, e.g., among 12 *Drosophila* species, to extensively study the evolution of genes and chromosomes on the *Drosophila* phylogeny and to discover and refine functional elements in *D. melanogaster* (Stark et al. 2007).

There is a growing body of evidence from comparative genomics analysis that a large amount of noncoding DNA sequences in *Drosophila* are under substantial selective

constraints, and may thus possess some potential functions. For example, Andolfatto (2005) has analyzed new and previously published polymorphism data for 35 coding fragments and 153 noncoding fragments scattered across the X chromosome of *D. melanogaster*, with a sample size of 12 *D. melanogaster* alleles and a single *D. simulans* sequence. He estimated that about 40%-70% of nucleotides in intergenic regions, UTRs and most intronic DNA are evolutionarily constrained relative to synonymous sites, and a substantial fraction of the nucleotide divergence in these regions was driven to fixation by positive selection. He then suggested that a large fraction of noncoding DNA is functionally important and subject to both purifying selection and adaptive evolution (Andolfatto 2005). A recent genome-wide analysis covering the whole euchromatic genome of *D. melanogaster* has indicated that functional constraints in noncoding *Drosophila* DNA are generally surprisingly high, >50% (Halligan and Keightley 2006). They also showed that nucleotide sites whose distance from the coding sequence boundary is up to 5kb are still under substantial selective constraints (Halligan and Keightley 2006). Both analyses used putatively neutrally evolving sites to estimate selective constraints, but different candidates, based on the neutral theory.

1.4. Transposable elements and their evolutionary importance

As discussed previously, TEs comprise a large fraction of noncoding DNA sequences and could be responsible for the C-value enigma. To unravel and refine the functional significance of noncoding DNA, studying the contribution of TEs to the host is an important issue. TEs are sequences of DNA that can move around to different positions

within the genome of a single cell, a process called transposition. They were first discovered by Barbara McClintock in the 1940s during her studies in maize (McClintock 1944, 1950 and 1951), but it has taken half a century to begin to understand how they act and how they affect the host genome.

TEs are widely distributed in bacteria, yeast, plants and animals, although their distributions vary dramatically among organisms. The long-term persistence of a TE family depends on the mobility rate of the existing TE copies (rate of new insertions), as well as the effects of new TE insertions on host fitness (Lynch 2007, p.167-168). There have been attempts to directly measure the rate of TE insertions using mutation accumulation experiments in *D. melanogaster* (Nuzhdin and Mackay 1995; Maside et al. 2000 and 2001). They estimated the average TE insertion rate to be $1.0\sim1.8\times10^{-4}$ per element per generation and the average excision rate to be $1.8\sim4.0\times10^{-6}$. If all insertions are neutral, this difference between these two rates will force the host to accumulate more TEs, of which the consequence is that the host genome will increase its genome size. However, most of new TE insertions appear to be deleterious and will not become fixed in the population. Furthermore, in *Drosophila*, the high intrinsic rate of DNA loss will also prevent TE insertions from pushing the genome to large size (Petrov et al. 1996).

Insertions that have slight/mild deleterious effects on host fitness are often of more interest, since they still have the ability to persist in the population for generations, or even become fixed. In a finite population, once such insertions occur in the host genome, the fate of this TE family in the population will depend on the rate of element giving

rising to new elements (daughter elements), the rate of TE loss by nonselective physical forces (that is associated with rate of DNA loss), and the fraction of new insertions that are effectively neutral (given the host population size) (Lynch 2007, p.174-177). Rate of new elements is positively correlated with the persistence of a TE family, while rate of TE loss is negatively correlated with the persistence. Because both of them are determined by intracellular activities, they should not be associated with any of the population properties, e.g., population size. The fraction of new insertions that are effectively neutral (and hence have the ability to become fixed in the population) is a function of population size and selection coefficient *s*. In general, for a small population, drift overwhelms selection; therefore, there will be a large fraction of new insertions that are effectively neutral to the host. Thus, this TE family may persist in the population for a long term. However, if the population size is large, drift will be a very slow process, and selection coefficient becomes the stronger determining factor for the fraction of fixable insertions. This TE family may become lost during the drift process, or eliminated from the population due to the relatively strong selection efficiency in a large population. Thus, the population size must be quite small for a TE family to establish or even become fixed in a population (also see Lynch 2007, p174-177). TEs that persist in a population for a long term, or even become fixed, may not only contribute to the genetic diversity of the organism via their mutational activities, but also may constitute a source of genetic innovation for the host acting as genes or gene regulatory elements (Biémont and Vieira 2006).

1.4.1. TE diversity

There are two main classes of TE: retrotransposons (class I elements), which are first transcribed into RNA and then reverse transcribed and reintegrated into the genome, and DNA transposons (class II elements), which are generally excised from one genomic site and integrated into another by a "cut and paste" mechanism. Retrotransposons are further subdivided into those that have "long terminal repeats" at their ends (LTR retrotransposons) and those that do not (non-LTR retrotransposons, also known as long interspersed nuclear elements, LINEs, in *Drosophila*). DNA elements are also known as terminal inverted repeat elements (TIR elements).

LTR retrotransposons contain *gag* and *pol* genes that encode a viral particle coat (GAG) and reverse transcriptase (RT), ribonuclease H (RH) and integrase (INT) to provide enzymatic activities for transpositions (Figure 1.2) (Kazazian 2004; Gregory 2005, p.171-175; Biémont and Vieira 2006). GAG specifies the activity of the RNA transposition intermediate of LTR retrotransposons. The RT enzyme is responsible for the synthesis of the double-stranded DNA (that will be integrated into the genome elsewhere) from a complementary single-stranded RNA, which is synthesized from the inserted DNA of the TE through the action of RNA polymerase II. The RH enzyme degrades the DNA-RNA hybrids obtained during transposition. The INT enzyme splices the double-stranded DNA into a new position in the host genome (Kazazian 2004; Biémont and Vieira 2006). Some LTR retrotransposons also contain envelope genes (*env*) that encode surface proteins that interact with the host cell membrane. However, LTR retrotransposons with *env* genes are only rarely able to move from one cell to another. Examples of TE moving from one cell to another are TEs from the *gypsy* family

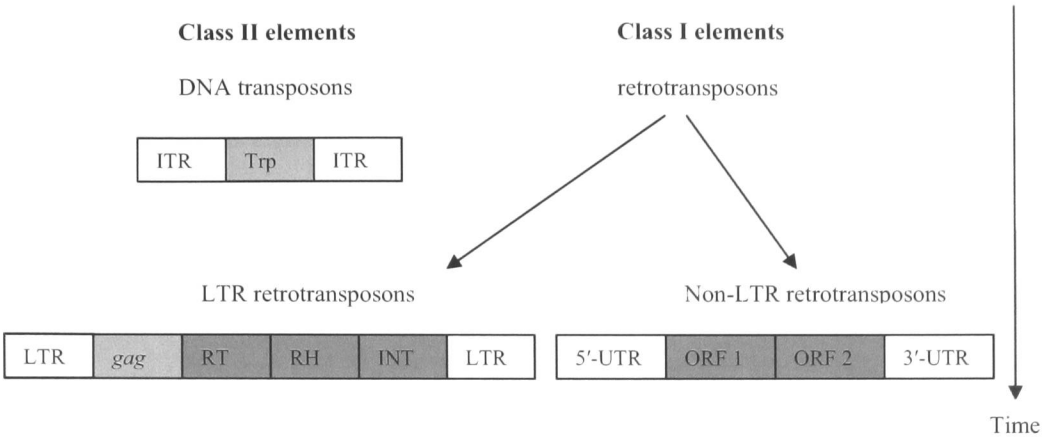


Figure 1.2. – Classes of mobile elements, DNA transposons, LTR retrotransposons, Non-LTR retrotransposons. The time scale is shown alongside. Class I elements are generally younger than class II elements. ITR stands for inverted terminal repeats; Trp stands for transposase; RT, RH and INT stand for reverse transcriptase, ribonuclease H and integrase, respectively, adapted from Figure 1 in Kazazian (2004) and Box 2 in Biémont and Vieira (2006).

(Kazazian 2004). Examples of LTR retrotransposons are the *Drosophila roo* and *gypsy* families, and the Ty3 element of *S. cerevisiae* (see Table 3.1 in Gregory 2005).

Non-LTR retrotransposons (LINEs) usually have two open reading frames (ORFs), one encoding a nucleic acid binding protein, and the other encoding an endonuclease and an RT enzyme (Figure 1.2). The 5'-UTR region contains an internal promoter (Kazazian 2004). Examples of non-LTR retrotransposons are Mammalian L1 elements, and R1 and R2 elements of *Drosophila* (Table 3.1 in Gregory 2005). Non-LTR retrotransposons transpose by simply reverse transcribing a complementary DNA (cDNA) copy of their RNA transcript directly onto the chromosomal target site (Gregory 2005, p.171-175). Because reverse transcription occurs on the target DNA after cleavage, the process is called target primed reverse transcription, or TPRT (Luan et al. 1993; Cost et al. 2002).

This process is very different from the process of reverse transcription of LTR retrotransposons, which occurs within the viral or virus-like particle in the cytoplasm (Voytas and Boeke 2002, p.631-662; Kazazian 2004), and is a complicated, multistep process (see Figure 2 in Kazazian 2004 and Figure 3.2 in Gregory 2005).

It is generally believed that class II elements are more ancient than class I elements (Figure 1.2) (Biémont and Vieira 2006). Class II elements move about the genome through a DNA intermediate, using the transposase (Trp) enzyme to split themselves in and out of the DNA (Figure 1.2) (Gregory 2005, p.176-178). Inverted terminal repeats (ITRs) are needed to facilitate the movement of DNA transposons. The transposase binds at or near the inverted repeats and to the target DNA, cutting the transposon from its old site and pasting into its new site in the genome through the breakage and joining reactions (Kazazian 2004; Gregory 2005, p.176-178). Examples of DNA transposons are the *Drosophila* INE-1 and P elements, and the fish Tc1/mariner elements (see Table 3.1 in Gregory 2005).

Note that the structure and function of eukaryotic mobile elements are in fact very similar to those of bacteria. It is then generally assumed that all eukaryotic TEs, either class I or II elements, are descended from bacterial elements (Eickbush and Malik 2001; Gregory 2005, p.179). This evolutionary scenario provides a rational basis for how TEs behave in the host, and also, how they co-evolve with the host in the long term.

1.4.2. The distribution of TEs within the genome

Many aspects of the transposition process are random, and in general, TE insertions occur in almost all genomic regions. However, it has been demonstrated that TEs are not randomly distributed within the genome (Gregory 2005, p193). For instance, some TEs are more often found in regions distant from host gene sequences, such as heterochromatin or regions between genes where recombination is reduced (Kidwell and Lisch 1997).

In the *D. melanogaster* genome, TE abundance tends to be higher in regions that lack recombination, but the correlation between element abundance and known differences in the recombination rate of various euchromatic regions is yet to be proved (Bartolomé et al. 2002; Rizzon et al. 2002; Eickbush and Furano 2002). In contrast, in the *C. elegans* genome, the distribution of retroelements is independent of the local recombination rate, and transposons tend to reside in regions of high recombination (Duret et al. 2000). This also suggests that patterns of TE distribution appear to differ greatly among different groups of organisms.

It has been suggested that the non-random distribution of TEs in the genome could be the consequence of three main factors: (1) some form of natural selection acting on TE insertions that have an impact on host fitness, (2) deleterious ectopic meiotic exchange between TEs of the same family, (3) TE target site specificities (Gregory 2005, p.193).

It has been strongly suggested that TE insertions are generally deleterious to the host, and will be eliminated by negative selection (Charlesworth and Charlesworth 1983; Kaplan and Brookfield 1983). Meanwhile, negative selection is also expected to act

against deleterious chromosome rearrangements induced by ectopic exchange between TEs of the same family inserted in nonhomologous locations (Langley et al. 1988; Montgomery et al. 1991; Charlesworth et al. 1992). These two mechanisms sometimes seem indistinguishable, since the selection efficacy is stronger in regions of high frequencies of recombination (Hill and Robertson 1966). Despite the mutagenic nature of TEs, some TEs do survive for variable lengths of time in coding regions or regions in which frequencies of recombination are high. This is because either these TEs have a neutral impact on host fitness and become fixed at random, or they benefit host fitness by taking part in gene regulation and are co-opted via positive selection. This has been supported by a growing body of evidence that TEs may have played important roles in host gene regulation and genome evolution in higher eukaryotic organisms (Jordan et al. 2003; van de Lagemaat et al. 2003; Kazazian 2004; Thornburg et al. 2006).

Within the euchromatic genome of higher organisms, TEs tend to accumulate in the pericentromeric and telomeric regions that lack recombination. For example, in the *D. melanogaster* genome, both the pericentromeric regions of the major chromosome arms and the entire chromosome 4 have higher densities of TE insertions, relative to non-pericentromeric regions (Kaminker et al. 2002; Rizzon et al. 2002; Bergman et al. 2006). Bartolomé et al. (2002) argue that this pattern is mainly due to the participation of TEs in ectopic recombination, which occurs when two TE copies inserted in nonhomologous positions in the same or different chromosomes may misalign because of their sequence homology. If exchange occurs, this may result in ectopic recombination (Gregory et al. 2005, p.194). Such ectopic recombination will result in deleterious chromosomal

rearrangements. Therefore, TEs in regions where recombination is reduced should be less deleterious, and hence more abundant. Note that there has been continuing debate about which of the two mechanisms, negative selection against deleterious TE insertion reducing host fitness or against deleterious ectopic exchange between TEs, is relatively more important in controlling TE abundance within the genome (e.g. Biémont et al. 1997; Charlesworth et al. 1997; Eickbush and Furano 2002).

In addition to effects of natural selection and ectopic recombination, it is suggested that the mechanism of "TE target site specificity" could also have impact on the non-random distribution of TEs within the genome (Gregory 2005, p.196-197). Although TEs can insert into many different locations in the genome, some TEs appear to have evolved strategies to minimize the potentially damaging effects of their inducing mutations on host fitness (Gregory 2005, p.196-197). These TEs prefer to insert into certain regions so that they would have little/no effect, or even some beneficial effect on host fitness. This is supported by the observation that *Drosophila* P elements exhibit a preference for inserting into a particular subset of genes as well as inserting near the 5′ end of gene transcription units within genes (Spradling et al. 1995; Gregory 2005, p.196-197). However, one should note that, although those elements (e.g., P elements) have insertion preference, their insertion preference are not such as to minimize the damage done by the elements. To distinguish the contribution of these three mechanisms on TE distributions is important, but often difficult in practice.

1.4.3. TEs in heterochromatin

Compared to the euchromatic portions of the genome, the highly condensed heterochromatic portions have relatively lower frequencies of recombination, and also lower gene density. In many eukaryotic species, TE elements tend to accumulate in heterochromatin because inserted elements are less likely to be deleterious, or less likely to cause chromosomal rearrangement (ectopic recombination). Indeed, it has been observed that TEs make up ~60% of the heterochromatic portions in both *D. melanogaster* and *Anopheles gambiae* genomes, while the proportion of TEs in the euchromatin is ~6% and 16% in these two species, respectively (Holt et al. 2002; Kapitonov and Jurka 2003). A concentration of TEs in heterochromatin is also seen in other species (e.g., *Arabidopsis thaliana*, *Tetraodon nigroviridis* and maize). It has been suggested that TE accumulation in heterochromatin was not caused by the long-term selection against euchromatic inserts (Dimitri and Junakovic 1999). Instead, heterochromatin may appear as a preferential target for (some) TEs, because TEs might be responsible for repairing DNA nicks whose density is high in heterochromatin (Labrador and Corces 1997; Dimitri 1997). Furthermore, TEs in heterochromatin may also have an impact on the evolution of heterochromatin, e.g., serving as regulatory sequences or promoting chromosomal rearrangements (Dimitri and Junakovic 1999). Thus, TE accumulation in heterochromatin is likely to reflect the evolutionary relationship between TEs and the euchromatic parts of the genome, rather than being mere addition of "junk DNA" to the genomic "wasteland" (Dimitri 1997; Dimitri and Junakovic 1999).

1.4.4. Possible domestication and application of TEs in the host genome

There has been continuing debate about the importance of TEs with respect to the host genome for the last four decades. TEs were once considered as just "selfish" genes or even junk for a long period of time. However, the releases of large-scale genomic sequences of different groups of organisms have shed light on the forces operating on repetitive sequences, and their potential roles in the host genome have gradually become clear. The mutagenic and parasitic characteristics of TEs may have enabled them to provide host genomes with the ability to generate new genetic diversity if necessary, and these so-called "junk DNAs" may play an important role in enhancing the evolutionary potential of their host (Kidwell and Lisch 2000; Lynch 2007, Chapter 7).

Firstly, TEs have the potential to contribute their own regulatory regions to form the host regulatory sequences, especially for those located in/near coding regions. Moreover, LTR retrotransposons usually carry relatively more regulatory signals than LINE and DNA elements, and may be more likely to be co-opted by the host (Fablet et al. 2006). A search of the Human Promoter Database has shown that ~25% of analyzed promoter regions contain a TE-derived sequence (Jordan et al. 2003), and TEs appear to serve as alternative promoters of many genes in human and mouse genomes (van de Lagemaat et al. 2003). TEs can also serve as enhancer elements for the host (Yang et al. 1998; Kidwell and Lisch 2000).

Secondly, TEs could contribute their coding potential to the host gene. For instance, Nekrutenko and Li (2001) examined 13,799 human genes and found 533 (~4%) cases of TEs within protein-coding regions. The majority of these TEs were first inserted within introns and were then recruited into coding regions as novel exons. A recent survey in

the *Bos taurus* genome has shown that ~2.37% of examined bovine genes contain TE-derived sequences within exons (Almeida et al. 2007). It is suggested that the association between TEs and exons may result from the novel alternative transcripts that have evolved a beneficial function, while the native transcript still maintains the original gene functions (DeBarry et al. 2005). In species where TEs are not as abundant as in mammals, it is still possible that TEs were inserted within coding regions and may have contributed their coding potential to the host gene (Ganko et al. 2006).

Thirdly, as mentioned above, TEs may play important parts in alternative splicing for the host gene. Alternative splicing provides an important mechanism for the host to generate the observed proteomic diversity from a relatively small number of protein-coding genes in eukaryotes (Gregory 2005, p.207). This mechanism allows the host to store the genomic information much more economically, and not to change the DNA content for the evolution of a new protein (Brett et al. 2001). Thus, the host will be able to adapt to new environments relatively faster, and produce a protein efficiently with improved functions under changing environmental conditions (Kreahling and Graveley 2004). Overall, TEs may have contributed to the host substantially, rather than being selfish or even junk.

1.4.5. Lineage-specific TE evolution and the effect of environments

It is suggested that TEs may be among the most lineage-specific elements of eukaryotic genomes. For example, a recent comparison among 12 vertebrate species indicated that the distribution of different TE types differed within and between these lineages, and

species-specific TE insertions accounted for the majority of size differences seen between lineages (Thomas et al. 2003). Even for close primate species, comparison between them revealed that transposition rates vary widely across lineages (Liu et al. 2003). In *Drosophila*, there are also great differences in TE distributions among closely related species. TEs in the two sibling species, *D. melanogaster* and *D. simulans*, differ considerably in amount and dynamics, with *D. simulans* having a smaller amount of TEs than *D. melanogaster* (Biémont and Cizeron 1999; Vieira and Biémont 2004). This observation may result from the demographic and geographical differences between the two species. Stresses due to changing environmental conditions and crosses between migrating populations could result in TE mobilization while a population colonizes (Vieira and Biémont 2004).

The influence of environments (e.g., climatic or trophic) could affect the evolution of organisms extensively. For instance, the action that the genome takes responding to the changing environments (e.g., methylation of certain DNA nucleotides, methylation or acetylation of the histone proteins) could create new genomic combinations that have better survival ability, increasing the variability of the genome, and allowing rapid evolutionary processes to take place within several generations (Arnault and Dufournel 1994). The behaviour of TEs under stress conditions is thus of particular interest, because these sequences are sources of mutations and therefore of genetic variability. In many organisms, TE sequences/fragments can control genes epigenetically when inserted within or close to them, by inducing genetic modifications including the methylation of nearby DNA or disruption of the normal epigenetic state of a nearby gene

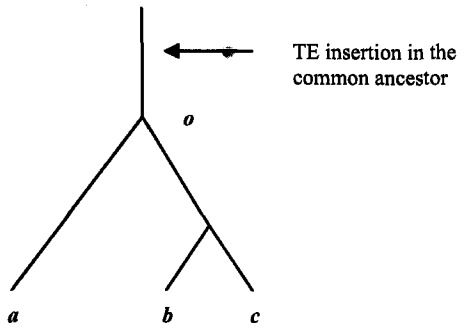


Figure 1.3. – TE insertion occurring in the common ancestor *o*, which leads to three sibling species *a*, *b* and *c*. An example of this is *D. melanogaster*, *D. simulans* and *D. sechellia*. Orthologous TEs among these three species are particularly of interest.

(Biémont and Vieira 2006). Thus, TEs may play an important role in population adaptation. As organisms tend to live in their preferred environments and have developed different ways to deal with changing environmental conditions, patterns of evolution of TEs and/or their interaction with hosts may differ (greatly) among organisms.

1.4.6. Analysis of orthologous TE elements between close species

When a TE was inserted in the common ancestor ahead of the speciation (Figure 1.3), there could be several consequences for this ancestral TE insertion: (1) this TE has become unfunctional sequences in the ancestral population since the insertion, undergoing structural decay and/or the process of coalescence. This TE may be transmitted to daughter species, or be deleted in the ancestral population; (2) this TE did get fixed in the ancestral population, and had undergone structural decay without being co-opted. Thus, fragments of this TE may still appear in the genomes of descendant

species, but they became diverged between species; (3) the ancient TE had become co-opted in the common ancestor or one/some descendant species (at the same time they may still undergo structural decay), and orthologous elements showed some sign of conservation among species. TEs under condition (3) are particularly of interest to us. Comparative genomics analysis can help reveal those elements conserved between closely related species.

1.5. Overview of the thesis

As an attempt to study the functional significance of noncoding DNA sequences (e.g., using transposable elements, TEs, as a group of candidates), this project mainly focuses on improving the global alignment quality for noncoding sequences and conducting comparative genomic analysis on TEs among closely related *Drosophila* species. One of the aims of the improved noncoding sequence alignment method is to develop more biologically realistic models consisting of probabilities for matches, mismatches and gaps (indels) of different lengths, and transition probability between matches/mismatches and indels. Such a method is able to automate noncoding sequence alignments solely based on explicit models for point substitutions and indel evolution without predefining any parameters. Meanwhile, this method should be able to find the most probable alignment with a reasonable speed, in that it can be used to carry out genomic analysis. This new method thus seeks for a good balance between alignment

quality and use of computer time, in particular when (noncoding) sequences under study are relatively long and/or distantly related.

Comparative genomic analysis on TEs in *Drosophila* is conducted based on alignments generated or refined by the improved noncoding sequence alignment method discussed above. The aims of such analysis include investigating forces operating on the evolution of TEs and their potential functions for the host, and detecting heterogeneities among TE classes and/or genomic regions in terms of the distribution within the genome and interspecies divergence. Such comparative analysis first focuses on finding putatively neutrally evolving sites (or the fastest evolving sites), and general forces operating on them. It then uses those fastest evolving sites as a neutral standard to investigate the functional signature of TEs of different classes or in different regions. Such analysis provides a powerful way to reveal the potential functionality of TEs in gene regulation and genome evolution.

Chapter 2 focuses on the development of the improved sequence alignment for noncoding sequences and benchmarking such a method along with other available methods. Chapter 3 focuses on investigating evolutionary patterns of candidates of putatively neutrally evolving sites, and Chapter 4 mainly focuses on revealing and refining the potential functional significance of some TEs. This will help understand more about the evolution of TEs contributing to the host. Chapter 5 discusses/concludes some of the main points presented in this thesis, and discusses some ideas for future work.

Chapter 2.

MCALIGN2: Faster, Accurate Global Pairwise Alignment of Non-coding DNA Sequences Based on Explicit Models of Indel Evolution

Jun Wang[§], Peter D. Keightley and Toby Johnson

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK.

[§]Corresponding author, Email: j.wang-13@sms.ed.ac.uk

JW, TJ and PDK all initially conceived of and co-designed the study. JW designed most of the statistical framework, developed, wrote and tested the software in this study, and wrote the manuscript. PDK participated in the design of the study, and commented on and wrote the manuscript. TJ guided the design of this study and provided the theoretical/technical support for the statistics/programming. The work of Equation (2.4), (2.5) and (2.6) was provided by TJ. TJ also provided JW the eigenvalue and eigenvector calculation c++ library for obtaining the instantaneous rate matrix. TJ commented on and wrote the manuscript as well. All authors read and approved the final manuscript.

2.1. Abstract

Background

Non-coding DNA sequences comprise a very large proportion of the total genomic content of mammals, most other vertebrates, many invertebrates, and most plants. Unraveling the functional significance of non-coding DNA depends on how well we are able to align non-coding DNA sequences. However, the alignment of non-coding DNA sequences is more difficult than aligning protein-coding sequences.

Results

Here we present an improved pair-hidden-Markov-Model (pair HMM) based method for performing global pairwise alignment of non-coding DNA sequences. The method uses an explicit model of indel length frequency distribution which can be specified, and allows any time reversible model of nucleotide substitution. The method uses a deterministic global optimiser to find the alignment with the highest posterior probability. We test MCALIGN2 in simulations, and compare it to a previous Monte Carlo based method (MCALIGN), to the pair HMM method of Knudsen and Miyamoto, and to a heuristic method (AVID) that performed very well in a previous simulation study. We show that the pair HMM methods have excellent performance for all combinations of parameter values we have considered. MCALIGN2 is up to ten times faster than MCALIGN. MCALIGN2 is more accurate in resolving indels given an accurate explicit model than heuristic methods, but is computationally slower.

Conclusions

MCALIGN2 produces better quality alignments by explicitly using biological knowledge about the indel length distribution and time reversible models of nucleotide substitution. As a result, it can outperform other available sequence alignment methods for the cases we have considered to align non-coding DNA sequences.

2.2. Background

The advent of automated DNA sequencing methods has resulted in an enormous growth in the volume of sequence data deposited in public databases. The increasing availability of genome sequence data for many related organisms offers great opportunities to study gene function and genome evolution, but it also presents new challenges for DNA sequence analysis, especially for non-coding DNA sequences.

For much of the past two decades, research in DNA sequence analysis has focused on protein-coding sequences, which account for only a very small proportion of the total genomic content in mammals, most other vertebrates, many invertebrates, and most plants (Li 1997). For example, protein-coding gene sequences comprise as little as 1-2% of the human and mouse genomes (International Human Genome Sequencing Consortium 2001; International Mouse Genome Sequencing Consortium 2002). However, there is an increasing body of evidence showing that non-coding DNA sequences contain many functional sequences involved in gene regulation and potentially other unknown functions. For example, it has been estimated that ~50% of bases in intergenic and intronic sequences of *Drosophila melanogaster* are selectively constrained (Halligan and Keightley 2006). In rodents, it has been inferred that the total number of selectively constrained nucleotides in non-coding DNA adjacent to gene sequences is similar to that in coding DNA (Keightley and Gaffney 2003). Evidence for the presence of a large number of potentially functional non-coding sequences on human chromosome 21 has recently been obtained from a comparative genomics analysis (Dermitzakis et al. 2002). Determining the fraction of non-coding DNA that is

functional and establishing what that function is, is therefore a central problem in genome research.

Accurate inferences about the function of non-coding DNA from comparative methods depend critically on correct alignments of non-coding sequences. However, the alignment of non-coding DNA sequences is more difficult than aligning protein-coding sequences. Protein-coding sequences tend to be highly evolutionarily conserved, so insertions and deletions (indels) are uncommon, and they usually occur in multiples of three base pairs. However, indel events are common in non-coding DNA, and can occur at most nucleotide sites. Numerous advances in sequence alignment methods for noncoding DNA have been made. Many recently proposed methods are based on heuristic alignment algorithms that can be very fast and accurate in cases where sequences are similar, but perform less well when sequence divergence is high (Pollard et al. 2004). Furthermore, heuristic scoring functions are not guaranteed to use the correct relationship between the relative penalties for point substitution and indel events, as they have no evolutionary interpretation. Therefore, explicit evolutionary models are desired to address this problem.

True evolutionary models of sequence evolution allow both multiple point substitutions and multiple indel events to affect any site in the sequence. The first true evolutionary model of indel evolution was introduced by Thorne, Kishino, and Felsenstein (1991), the TKF91 model, and allows single-residue indel events. This method uses a maximum likelihood algorithm to estimate the evolutionary distance between two sequences, summing over all possible alignments in the likelihood calculations (Thorne et al. 1991).

It was subsequently improved by allowing longer indel events with a geometric length distribution (Thorne et al. 1992), by assuming that the sequence contains unbreakable fragments, and that only whole fragments are inserted and deleted. This assumption introduces hidden information in the form of fragment boundaries, and may potentially bias multiple alignment (Miklos and Toroczkai 2001). Knudsen and Miyamoto (2003) presented a pairwise statistical alignment method based on an explicit evolutionary model of indel events. Indel length was assumed to be geometrically distributed, and up to two overlapping events were allowed for indels. A good approximation to such a model was then made using a pair HMM. The geometric distribution parameter, the indel rate, and the evolutionary time were estimated by maximum likelihood. A "long indel" evolutionary model has been introduced recently by Miklos et al. (2004), which allows multiple-residue indels without hidden information such as fragment boundaries. They developed a finite trajectory approximation for computing the likelihood function, producing a method that has very good performance (Miklos et al. 2004).

Previously, Keightley and Johnson (2004) proposed a non-coding sequence alignment method called MCALIGN. This is based on a simplified evolutionary model that does not allow for any multiple hits or interaction between indel events. A key feature of their approach is that it uses additional data from "unambiguous" alignments (e.g. between sequences from closely related species) to infer the actual distribution of indel lengths, and the relative rate of indels to point substitutions. They used a Monte Carlo (MC) hill-climbing algorithm to search for the most probable alignments. This method has been successfully used for aligning real genomic sequences, such as *Drosophila*, rodent and

hominid non-coding DNA (Keightley and Gaffney 2003; Haddrill et al. 2005; Keightley et al. 2005). In a simulation study, Keightley and Johnson (2004) found that MCALIGN was generally superior to the other alignment methods that it was compared to.

Here, we describe an improved non-coding sequence alignment algorithm based on a generalisation of the evolutionary model used by Keightley and Johnson (2004). We show how a combination of a dynamic programming (DP) algorithm and a one dimensional deterministic optimisation-algorithm can be used to find the most probable pairwise sequence alignment. Note that when we assume the Jukes-Cantor model for nucleotide substitution (Jukes and Cantor 1969), the present DP method and the previous MC method are essentially using two different optimisers to attempt to maximise the same "score" function: alignment probability. However, the new optimiser is expected to be better and faster.

We have compared our method to the pair HMM method of Knudsen and Miyamoto (PairHMM_KM hereafter), which is quite similar to the present method in that it explicitly makes use of an evolutionary time parameter (Knudsen and Miyamoto 2003). We have also compared our method to the heuristic alignment program AVID of Bray et al. (2003) in simulations that assume a general-time-reversible (GTR) model (Lanave et al. 1984) that had first been fitted to real *Drosophila* non-coding DNA sequence data. It has been shown that AVID performs very well compared to other heuristic methods (Keightley and Johnson 2004, Bray et al. 2003), so here we only compare our method to AVID rather than other heuristic methods.

In our tests, the new DP method (MCALIGN2) is up to ten times faster than the previous MC method (MCALIGN), and is also faster than the pairHMM_KM method (Knudsen and Miyamoto 2003), although none can compete in speed terms with heuristic methods.

For cases of real non-coding sequence data, we also compared MCALIGN2 with AVID and CLUSTALW (Thompson et al. 1994), and show that they perform differently for some specific cases.

2.3. Implementation

We use a Bayesian statistical framework (Gelman et al. 2003; Durbin et al. 1998) to make inference about the pairwise alignment. The aim is to compute the posterior probabilities of different possible alignments, using the observed sequences as data and eliminating other "nuisance" parameters from the analysis. Here we focus on finding the alignment with the highest posterior probability.

Let t be the total divergence time between two sequences, a be an alignment of two sequences, and S be the observed data, which is two non-coding DNA sequences. In a Bayesian framework, the behaviours of all variables are modelled by probability distributions. Joint inference about a and t is accomplished simply via Bayes' theorem

$$P(a,t \mid S) = P(a,S \mid t)P(t)\frac{1}{P(S)}. \qquad (2.1)$$

The probability $P(S)$ that appears in the denominator of Equation (2.1) may be difficult to calculate, but because in Bayesian inference the observed data S is held fixed, the unconditional probability $P(S)$ is constant. We can therefore make our inference using only relative probabilities and $P(S)$ need not be calculated. The other unconditional probability that appears in Equation (2.1) is $P(t)$, which is specified as a prior; our method will work for any prior.

To calculate the posterior probability of an alignment, we consider the divergence time t as a nuisance parameter. The posterior probability for an alignment is therefore marginal to the divergence time t, and is calculated using the integral

$$P(a \mid S) = \int P(a,t \mid S)dt. \tag{2.2}$$

We approximate this integral using Laplace's method, described in detail below.

Probability Model of Sequence Evolution

The most difficult probability to specify in Equation (2.1) is $P(a,S|t)$, which is the joint probability of alignment a and sequences S given a divergence time t. This probability is specified according to a model. Here, we use the pair hidden Markov model (HMM) shown in Figure 2.1. For a comprehensive introduction to pair HMMs, see the books by Durbin et al. (1998) and Ewens and Grant (2001). For a given time t, the pair HMM shown in Figure 2.1 generates the sequence alignment by using a series of transitions between states, accompanied by emissions. Once in a given state, the transition probabilities (shown in Figure 2.1) govern which state the pair HMM will move to next.

Upon arrival at a new state, the pair HMM emits some observed data according to the emission probability distributions (shown in Figure 2.1). For example, state M has emission probability distribution p_{minj} for emitting an aligned base pair m_i:n_j, and state I_x and I_y have distributions q_{mi} and q_{nj} for emitting nucleotide base m_i and n_j against a gap, in each of the two sequences (labelled x and y respectively).

The transition probabilities for the pair HMM determine the pattern of indels in the alignment. The emission probabilities for the pair HMM determine the sequences that are observed, given the pattern of indels in the alignment. We specify the transition probabilities with an explicit model of insertion and deletion events, and the emission probabilities are specified by a model of nucleotide frequencies and of nucleotide substitutions. We consider the transition and emission probabilities in turn.

We assume that insertions and deletions occur as independent events over time with a total rate θ per interbase site relative to nucleotide substitutions. As we ignore multiple hits for indels, the probability of an indel is $1\text{-}e^{(-\theta t)}$ per interbase site, which we approximated as θt, an approximation that should be good for small values of t. An indel can correspond to a gap in sequence x or a gap in sequence y. These two events have the same probability, so the probability of a gap in either of the two sequences, x and y, is then $\theta t/2$. In Figure 2.1, this corresponds to the transition probability from the M state to the $I_{x,1}$ state, or to the $I_{y,1}$ state. The pair HMM must move through one of these states whatever the length of the indel.

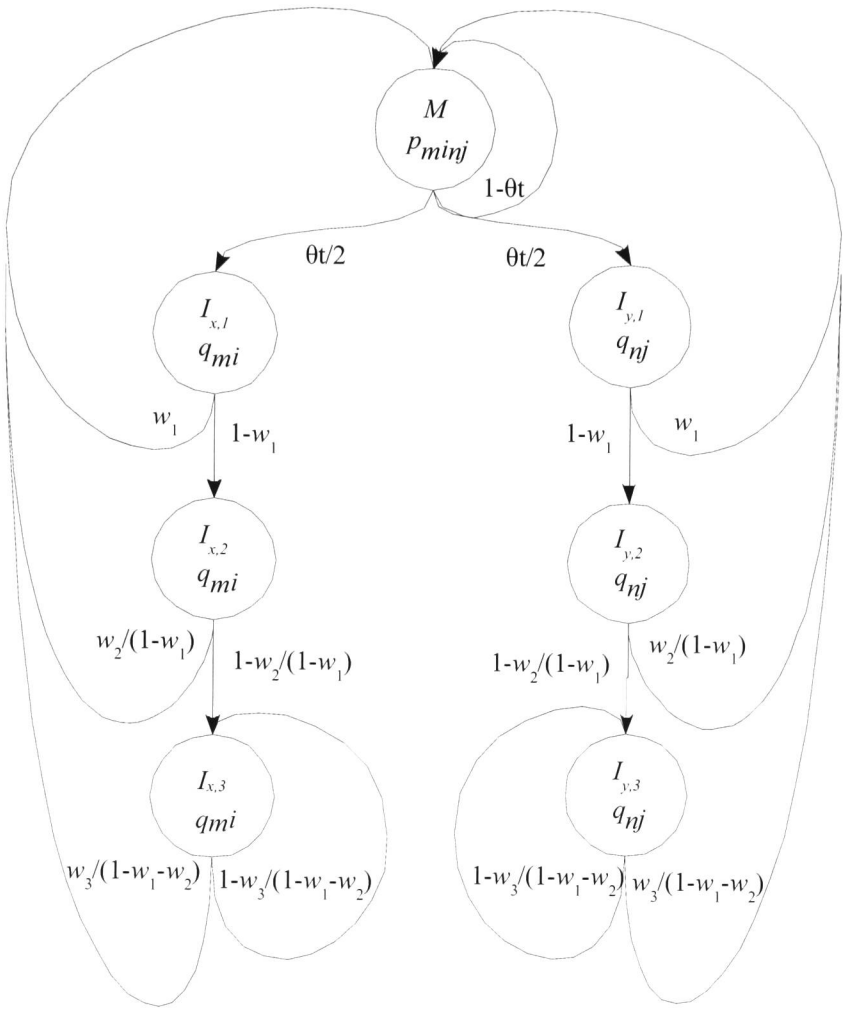


Figure 2.1. – Pair HMMs assuming an affine gap model. Assume two homologous sequences x and y. Let m_i be the ith nucleotide in sequence x and n_j be the jth nucleotide in sequence y. M represents the state that m_i is aligned to n_j, I_x represents the state that m_i is aligned to a gap (in an insertion with respect to y), and I_y represents the state that n_j is in an insertion with respect to x. The numbers shown after x or y indicate the positions of m_i and n_j in the insertion with respect to the other sequence. The transition probability is shown between states.

The standard affine gap model corresponds to assuming that the lengths of indels follow a geometric distribution (Durbin et al. 1998; Lunter et al. 2004). Empirical data

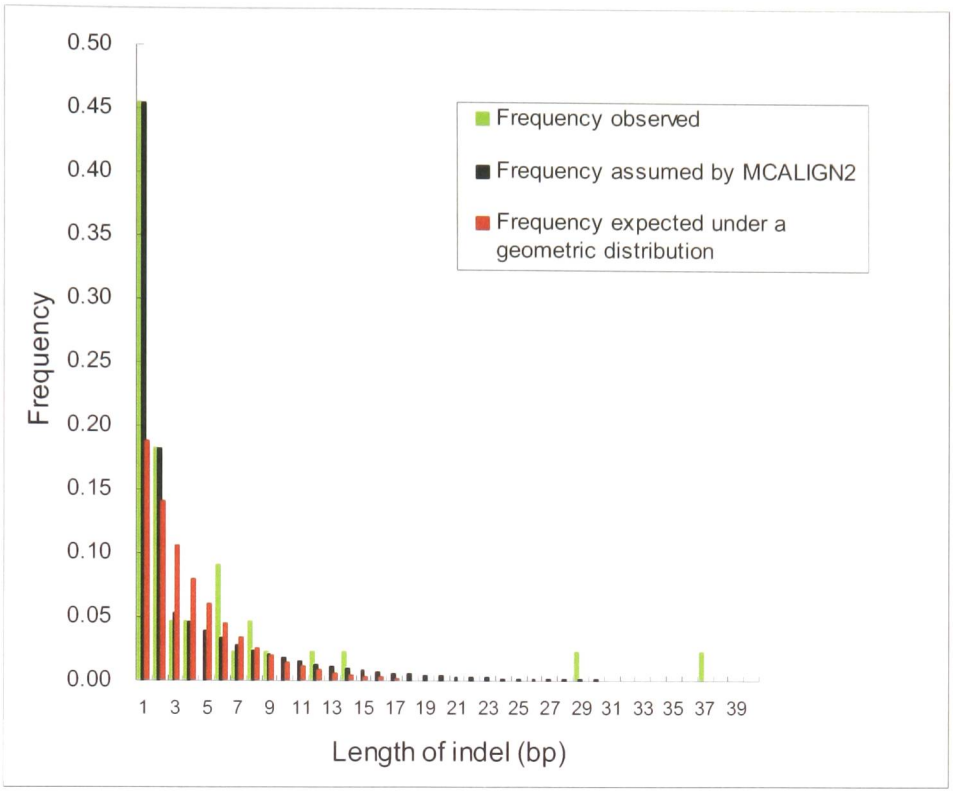


Figure 2.2. – The empirical distribution of indel lengths in noncoding DNA between *D.simulans* and *D.sechellia* from Keightley and Johnson 2004 (green histogram), the indel length frequency distribution assumed by MCALIGN2 (black histogram) and the indel length frequency distribution expected under a geometric distribution based on maximum likelihood estimation given the observed data (red histogram).

on indel lengths in *Drosophila* non-coding DNA show an obvious departure from a geometric distribution, since 1- or 2-residue indels are more common than expected (Figure 2.2). Therefore, our model includes separate parameters for the probabilities of indels of length 1-bp and 2-bp, since these can be reliably estimated. Because there are less data on the length distribution for longer indels, we assumed a geometric distribution. Other more complex distributions are widely preferred for protein sequence

alignments (Miller and Myers 1988; Miklos et al. 2004), but their large numbers of parameters cannot be reliably fitted using available data for noncoding sequences. Let w_i be the probability of an indel of length i, with $w_{i+2}/w_{i+1}=w_{i+1}/w_i$ for $i\geq 3$, and $\sum_i w_i=1$. We follow the approach of Keightley and Johnson (2004) and estimate these parameters, along with the parameter θ that describes the total rate of indels relative to nucleotide substitutions, from additional data, as described below.

As shown in Figure 2.1, given that the pair HMM has arrived in state $I_{x,i}$ or $I_{y,i}$ (for any $i \geqslant 1$), the transition probability back to the M state is $w_i/(1-\sum_{j=1}^{i-1} w_j)$ and the transition probability to state $I_{x,i+1}$ or $I_{y,i+1}$ is $1-w_i/(1-\sum_{j=1}^{i-1} w_j)$. (Here a sum with no terms is understood to be zero.) This produces the desired distribution of indel lengths. We also assume that a gap in sequence x will not be followed directly by a gap in sequence y, and therefore there are no transitions from any of the states I_x to any of the states I_y, or vice versa. Our approach could be extended to accommodate an indel length distribution that is any mixture of geometric distributions (as used by Miklos et al. (2004)) by duplicating the nodes in the pair HMM for insertions and deletions, and setting transition probabilities according to each component in the mixture. Such an extension may lead to increased accuracy, but at the expense of increased computational demands.

In order to make our pair HMMs describe a probability distribution over all possible alignments, we need to include a Begin state and an End state. We set the transition probability from the Begin state to states M, $I_{x,1}$ and $I_{y,1}$ to be the same as those from the M state. We allow all states to make transitions to the End state, with a low transition

probability ε. If ε is small enough, we can ignore it in all of our calculations (Durbin et al. 1998).

The emission probabilities, which determine the sequences given the pattern of indels, are derived from the general time-reversible (GTR) model of nucleotide substitution (Lanave et al. 1984). The Jukes-Cantor model (Jukes and Cantor 1969) and the Kimura-2-parameter model (Kimura 1980) are two specific cases of the GTR model when certain parameters are fixed.

The emission probabilities q_{mi} and q_{nj} are the equilibrium frequencies of nucleotides m_i and n_j, which are equal for sequences x and y. The emission probabilities p_{minj} are the probabilities of starting with an unobserved common ancestor nucleotide o, drawn from the equilibrium distribution of nucleotide frequencies, and evolving independently down two lineages, to m_i in time t_1 along one lineage and to n_j in time t_2 along the other lineage. (Under a time reversible model, this is the same as the probability of starting with n_j and evolving to m_i (or vice versa) in time t_1+t_2.) Since the times t_1 and t_2 are individually nonindentifiable, we parameterise our model simply by the total divergence time $t = t_1+t_2$. For a given total divergence time, the conditional probability of evolving to m_i given starting with n_j is $p_{mi|nj} = p_{minj}/q_{nj}$, and the matrix of these conditional probabilities, $Q(t)$, can be calculated from the fixed instantaneous rate matrix A by matrix exponentiation (Felsenstein 2004), that is,

$$Q(t) = e^{tA}. \tag{2.3}$$

which can be calculated using the eigenvalues and eigenvectors of A. Here we estimate the rate matrix A from the same external data that is used to estimate the parameters for indels, as described below.

Alignment Algorithm

Given that $P(a,S|t)$ has been specified by the model, and that a prior $P(t)$ for the divergence time has also been specified, we have developed an algorithm to infer the approximate maximum a-posteriori (MAP) alignment $\hat{a}$. This is the alignment with highest posterior probability given the observed sequences, with the divergence time eliminated as a nuisance parameter. Thus, $\hat{a}$ is the alignment that maximises $P(a|S)$, which is given by the integral in Equation (2.2). To approximate this integral, we assume that $P(a,t|S)$, when treated as a function of t with both a and S held fixed, is approximately Gaussian. Then, using Laplace's method (O'Hagan and Forster 2004), we can write

$$P(a \mid S) \approx P(a, \hat{t}_a \mid S)\sqrt{2\pi}\,|V_a|^{\frac{1}{2}}. \tag{2.4}$$

Here, $\hat{t}_a$ is the mode, or value of t that maximises $P(a,t|S)$ (again, when treated as a function of t with a and S held fixed). The quantity

$$|V_a| = \frac{1}{\dfrac{d^2 \ln P(a,t \mid S)}{dt^2} \Big| t = t_a} \tag{2.5}$$

is the modal dispersion, which is the reciprocal of the curvature at the mode . We make a further approximation,

$$P(a \mid S) \approx P(a, \hat{t}_a \mid S) \times C(S) \tag{2.6}$$

where $C(S)$ is a constant that depends on S but not on a or t. This approximation can be made when we wish to maximise $P(a|S)$ over a set of a for which $|V_a|$ is approximately constant. The goodness of this approximation is discussed below.

Given that our approximations hold, Equation (2.6) shows that $\hat{a}$ maximises $P(a|S)$ if and only if $\hat{a}$ maximises $P(a, \hat{t}_a \mid S)$. Since by definition $(\hat{a}, \hat{t}_a)$ maximises $P(a,t|S)$, we see that $\hat{a}$ can be found by unrestricted optimisation of $P(a,t|S)$. Our algorithm to find $\hat{a}$ exploits the fact that we are free to solve the unrestricted optimisation problem with any manner we choose, and specifically that we can "change the order of maximisation". The statistical argument we presented above says that we should find $\hat{t}_a$ for each a, and then maximise $P(a, \hat{t}_a \mid S)$ over all a. An equivalent solution is to find $\hat{a}_t$ for each t (that is, the best alignment for a given t) and then maximise $P(\hat{a}_t, t \mid S)$ over all t. The second solution is much easier in practice, because $\hat{a}_t$ can be found using a standard dynamic programming algorithm for pair HMMs (Durbin et al. 1998), and then $P(\hat{a}_t, t \mid S)$ can be maximised using any standard algorithm for maximising a one dimensional function.

The dynamic programming algorithm guarantees to find the global maximising $\hat{a}_t$ (with ties broken arbitrarily). We find a straightforward Golden Section Search (Press et al.

1992) to be adequate for maximising $P(\hat{a}_t, t \mid S)$. This assumes that there is a single global optimum to be found. Actually we are able to trap events where local optima are detected. However, no local optima have ever been detected. We terminate the search when the values of t bracketing the maximum differ by less than 0.001. Moreover, we are able to terminate earlier when the optimal alignment is the same at all points within the bracketing area.

Parameterization of Models of Sequence Evolution

Our model of noncoding DNA evolution is parameterized according to the empirical distribution of indel lengths and their overall rate relative to nucleotide substitutions from species for which essentially unambiguous alignments can be made. Here, we consider a parameterization by intronic data of *D. simulans* and *D. sechellia* (Shown in Figure 2.2). For these data, the rate of indels per interbase site, relative to the rate of nucleotide substitution, was previously estimated as $\theta = 0.225$ (Keightley and Johnson 2004). We fitted the observed frequencies of different indels lengths to our model as follows. We directly use the observed frequencies of 1-bp and 2-bp indels, that is, 0.455 and 0.182, respectively. For indels of ≥3-bp, the frequencies, W_x, for the model were obtained by minimizing the sum over ≥3-bp indels of the squared differences between the observed frequency distribution and $w_x = \beta/\alpha^x$. Here β is a constant. The estimate for α was 1.170. Our software performs this curve fitting and in fact the whole analysis with a supplied empirical distribution containing any lengths.

A GTR model of nucleotide substitution was fitted to *Drosophila* data shown in Table

Table 2.1. *Drosophila* intronic data that is used to derive a GTR model of DNA evolution.

		Sequence 2				
		A	G	C	T	total
Sequence 1	A	1363	45	18	54	1480
	G	37	823	9	17	886
	C	21	11	898	32	962
	T	17	11	27	1120	1175
	total	1438	890	952	1223	4503

Pairs of nucleotide for 4503 sites of sequence that has diverged according to a general-time-reversible (GTR) model, from real *Drosophila* intronic data. The columns are the bases in the first sequence. Here, we chose a long intron from *D. simulans* and *D. melanogaster*, aligned them using AVID, then counted the aligned sites regardless of gaps.

2.1. By assuming the GTR model, we can then symmetrise this matrix by averaging the table with its transpose before any of the following calculations were carried out. The estimated equilibrium frequencies of each base are obtained from the normalised column sums, yielding $(q_A, q_G, q_C, q_T) = (0.324, 0.197, 0.213, 0.266)$. The estimated rates of each type of substitution are obtained by dividing the entries in each column by the respective column sums, yielding:

$$\hat{Q} = \begin{Bmatrix} 0.934201 & 0.0461712 & 0.0203762 & 0.029608 \\ 0.0281014 & 0.926802 & 0.0104493 & 0.0116764 \\ 0.0133653 & 0.0112613 & 0.938349 & 0.0246038 \\ 0.0243317 & 0.0157658 & 0.0308255 & 0.934112 \end{Bmatrix}. \tag{2.7}$$

Finally, find the matrix A that satisfies Equation (2.3) when time is measured in units of expected substitutions, to obtain our estimate of the instantaneous rate matrix:

$$\hat{A} = \begin{Bmatrix} -0.995107 & 0.706988 & 0.301277 & 0.446817 \\ 0.430299 & -1.1037 & 0.153414 & 0.17136 \\ 0.197616 & 0.165335 & -0.922152 & 0.373111 \\ 0.367192 & 0.231375 & 0.467461 & -0.991288 \end{Bmatrix}. \quad (2.8)$$

Performance Evaluation

For non-coding sequences, there are few externally verified alignments available to test the performance of alignment methods. As a substitute, we simulate sequence divergence *in silico*, so that sequences are generated that are related by a known, "correct" alignment (Pollard et al. 2004). We tested the MCALIGN2 program by examining the posterior probability of the best alignment found by the algorithm, the fraction of correctly aligned sites, an estimate of divergence time calculated from the estimated alignment, and the time taken to compute the alignment.

We compared the dynamic programming approach used here against the Monte-Carlo approach proposed previously (Keightley and Johnson 2004) and the pair HMM approach of Knudsen and Miyamoto (2003) in simulations assuming the Jukes-Cantor model of nucleotide evolution. In comparisons of MCALIGN2 and MCALIGN, for each simulated pair of sequences, we compared the posterior probability, $P(a|S) \simeq P(a, \hat{t}_a \mid S)$, of the best alignment found by MCALIGN2 with the best alignments found by MCALIGN.

We also compared MCALIGN2 against AVID of Bray et al. (2003) in simulations assuming a GTR model, parameterised using the *Drosophila* intronic data as described

above. In these comparisons we investigated cases in which the model assumed by MCALIGN2 differed from the simulation model, by using the simpler JC and K2P models to analyse data simulated under a GTR model.

In all comparisons, we calculated the fraction of correctly aligned sites by counting the number of base pairs or bases-to-gaps which were correctly aligned in a comparison to the true alignment. As an alternative measure of alignment quality, we considered the precision of divergence time estimated from the alignments. The estimator of divergence time we used was distance under the GTR model. It is made by estimating the base frequencies q_i, and the rates a_{ij}, and finding ones that most closely predict the observed net transition matrix P (Felsenstein 2004). This estimator of divergence time uses only the non-indel regions, and does not use the presence of indels to help estimate divergence time. For all the simulations with a given divergence time t and a certain evolutionary model, we observed the mean and variance of the estimator of t calculated from both the true alignment and the alignments estimated by sequence alignment methods we considered here. We express the precision of the estimator as the estimated root mean squared error (e.r.m.s.e.), since none of the estimators examined are perfectly unbiased. For t, this is

$$e.r.m.s.e. = \sqrt{\frac{1}{N}\sum (t_{est} - t_{true})^2} \qquad (2.9)$$

when there are N simulations.

Although our program allows any prior for divergence time, for all comparisons we used a relatively diffuse or uninformative prior:

$$P(t) = \frac{4}{3} e^{(-\frac{4}{3}t)} \qquad (2.10)$$

which has the mean 0.75. Because low divergences are more likely than high ones for two homologous sequences, this prior on t seems to be a reasonable one.

2.4. Results

Comparison amongst PairHMM_KM, MCALIGN2(DP) and MCALIGN(MC)

We generated non-coding sequence data using a model of non-coding DNA evolution in which gap lengths are parametrized by intronic data of *D. simulans* and *D. sechellia*, and point substitutions occur according to the Jukes-Cantor model to compare the performances of PairHMM_KM, MCALIGN2 (DP hereafter) and MCALIGN (MC hereafter). In this setting, the DP and MC methods aim to find the same most probable alignment, since they assume essentially the same model and prior, but use different algorithms.

Table 2.2 and 2.3 show the mean and e.r.m.s.e. of estimated divergence time (t), and the proportions of correctly aligned sites for combinations of θ and t. All alignment methods perform similarly when the true divergence time is not too great, $t \leq 0.2$, and

Table 2.2. Performance of MCALIGN2(DP), MCALIGN(MC) and PairHMM_KM compared by the estimator of divergent time corrected by the Jukes-Cantor model.

Simulated			**Alignment Estimated**		
t	θ	Alignment Known	PairHMM_KM	MC	DP
0.05	0.225	0.0502 (0.0107)	0.0496 (0.0103)	0.0499 (0.0104)	0.0501 (0.0103)
0.10	0.225	0.0998 (0.0146)	0.0987 (0.0152)	0.0994 (0.0153)	0.0989 (0.0154)
0.15	0.225	0.1493 (0.0208)	0.1482 (0.0226)	0.1507 (0.0230)	0.1487 (0.0208)
0.20	0.225	0.2025 (0.0241)	0.1994 (0.0267)	0.2053 (0.0263)	0.2001 (0.0256)
0.25	0.225	0.2515 (0.0286)	0.2440 (0.0348)	0.2593 (0.0346)	0.2476 (0.0319)
0.30	0.225	0.3003 (0.0311)	0.2955 (0.0419)	0.3162 (0.0525)	0.2981 (0.0349)
0.15	0.10	0.1519 (0.0198)	0.1503 (0.0189)	0.1502 (0.0189)	0.1500 (0.0188)
0.15	0.30	0.1515 (0.0202)	0.1507 (0.0226)	0.1566 (0.0234)	0.1523 (0.0218)
0.15	0.40	0.1512 (0.0194)	0.1480 (0.0220)	0.1645 (0.0263)	0.1516 (0.0213)

Estimates of sequence divergence, t, from 200 replicates for each combination of t and θ, with sequences of length 500 base pairs. Estimated root mean square error (e.r.m.s.e.) is shown after divergence time in parentheses.

the indel rate is not too great, $\theta \leq 0.3$. For these parameters, the fraction of correctly aligned bases is greater than 90% and is similar for all the three methods. The mean estimated divergence time calculated from estimated alignments is close to the true values, and the e.r.m.s.e. is not substantially greater than if the true alignment is known. However, when the divergence time t became larger ($t > 0.2$) or the indel rate becomes larger ($\theta = 0.4$), the performance of the MC method becomes noticeably inferior, since the mean proportion of correctly aligned bases is significantly lower than for the alignments estimated by the DP and PairHMM_KM method, and the divergence time estimates are more biased and have larger e.r.m.s.e.. For the largest indel ratio we considered, $\theta = 0.4$, the MC method tends to estimate an alignment with many gaps and

Table 2.3. Performance of MCALIGN2(DP), MCALIGN(MC) and PairHMM_KM compared by examining the proportions of correctly aligned sites.

Simulated		Proportion of correctly aligned sites		
t	θ	PairHMM_KM	MC	DP
0.05	0.225	0.992 (0.0072)	0.992 (0.0078)	0.993 (0.0074)
0.10	0.225	0.977 (0.0133)	0.977 (0.0137)	0.977 (0.0127)
0.15	0.225	0.954 (0.0210)	0.951 (0.0235)	0.955 (0.0186)
0.20	0.225	0.920 (0.0300)	0.915 (0.0345)	0.922 (0.0293)
0.25	0.225	0.868 (0.0416)	0.850 (0.0596)	0.869 (0.0433)
0.30	0.225	0.810 (0.0500)	0.761 (0.0863)	0.813 (0.0511)
0.15	0.10	0.984 (0.0116)	0.982 (0.0102)	0.983 (0.0108)
0.15	0.30	0.933 (0.0224)	0.925 (0.0292)	0.933 (0.0246)
0.15	0.40	0.905 (0.0325)	0.894 (0.0376)	0.906 (0.0329)

Proportion of matched bases from 200 replicates for each combination of t and θ, with sequences of length 500 base pairs. Standard deviation of mean is shown after the proportion of matched bases in parentheses.

the estimates of t tend to be higher than the true values. Table 2.3 also shows that the DP and PairHMM_KM methods both have more stable performances for most of the cases we have considered, in the sense of producing lower standard deviations of proportions of correctly aligned sites. It is also shown that the efficiency of MCALIGN2 is generally slightly better than PairHMM_KM.

For the same simulated datasets, Figure 2.3 compares the log values of alignment probability for MCALIGN2 and MCALIGN, since they use essentially the same scoring function. For the two methods, the approximation of Equation (2.6) was used to calculate alignment probability marginal to divergence time. Both methods perform equivalently for almost all the simulations when divergence time is very small (t = 0.05);

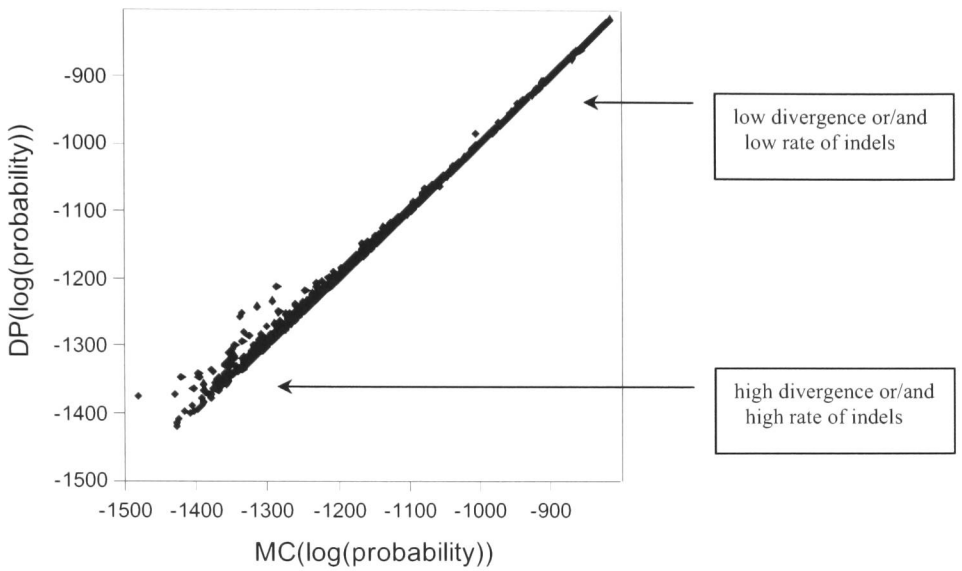


Figure 2.3. – Probability test using the probability function of MCALIGN2(DP), comparing the performances of alignments produced by the DP method and the MC method. All the values here are log values.

we presume that both methods are able to find the globally most probable alignment. However, when divergence time and/or rate of indel events becomes larger, the DP method begins to outperform the MC method, in the sense that the alignments produced by MCALIGN2 have higher probabilities. For the highest divergence time (t = 0.30) and/or rate of indel events (θ = 0.40) we considered, the DP method outperformed the MC method for almost all of the replicate simulations. This clearly indicates that the MC algorithm of Keightley and Johnson (2004) gets stuck at local optima.

Comparison between MCALIGN2 and AVID

For each combination of values of t and θ, 200 replicate simulations were performed, each simulating a pair of sequences of length 500 base pairs, evolving under an indel model and a general time reversible (GTR) model of nucleotide substitution, parameterised using real Drosophila data. This model is very different from the simple Jukes-Cantor (JC) model, and quite different to Kimura's 2 parameter (K2P) model. In addition to comparing MCALIGN2 with AVID, it is interesting to explore the effect of the nucleotide substitution model assumed by MCALIGN2. We aligned each simulated pair of sequences using MCALIGN2 under the assumptions of the correct GTR model, a simple K2P model with the ratio of transition events to transversion events equal to 2, and the JC model.

The results in Table 2.4 and 2.5 show that the alignments found by MCALIGN2, when the correct GTR model was assumed, are more accurate for almost all combinations of parameter values we have considered. In comparison, alignments found by MCALIGN2, when the incorrect JC or K2P models were assumed, are only slightly less accurate. Alignments found by AVID generally have the lowest accuracy in the cases studied.

Here, lower accuracy is indicated by a lower proportion of correctly aligned bases, and estimates of divergence time t that are more biased and have larger e.r.m.s.e.. In particular, alignments produced by AVID exhibit consistent upward bias estimates of t, and lower mean proportions of correctly aligned bases than alignments produced by MCALIGN2. This remains true, for most of the cases we considered here, whether MCALIGN2 used the correct GTR model of nucleotide substitution, or the incorrect JC

Table 2.4. Performance of MCALIGN2 and AVID compared by proportions of correctly aligned bases on a GTR model.

Simulated		Proportion of matched bases			
t	θ	AVID	MCALIGN2(JC)	MCALIGN2(K2P)	MCALIGN2(GTR)
0.05	0.225	0.991 (0.0085)	0.993 (0.0057)	0.993 (0.0057)	0.993 (0.0057)
0.10	0.225	0.973 (0.0141)	0.978 (0.0127)	0.979 (0.0127)	0.979 (0.0127)
0.15	0.225	0.946 (0.0283)	0.954 (0.0212)	0.956 (0.0184)	0.958 (0.0184)
0.20	0.225	0.904 (0.0325)	0.916 (0.0283)	0.920 (0.0269)	0.922 (0.0269)
0.25	0.225	0.852 (0.0452)	0.867 (0.0438)	0.873 (0.0410)	0.876 (0.0396)
0.30	0.225	0.795 (0.0566)	0.811 (0.0495)	0.824 (0.0481)	0.831 (0.0481)
0.15	0.10	0.980 (0.0141)	0.982 (0.0113)	0.982 (0.0113)	0.983 (0.0113)
0.15	0.30	0.913 (0.0283)	0.935 (0.0226)	0.936 (0.0212)	0.941 (0.0212)
0.15	0.40	0.876 (0.0354)	0.900 (0.0311)	0.905 (0.0297)	0.916 (0.0283)

Proportion of matched bases from 200 replicates for each combination of t and θ, with sequences of length 500 base pairs. Standard deviation of mean is shown after the proportion of matched bases in parentheses. Here MCALIGN2 is tested by assuming either the correct model of DNA evolution (GTR) or the incorrect models (JC and K2P).

or K2P models. The improvement in alignment quality gained by knowing the correct model of nucleotide substitution is generally modest, but worthwhile.

Test using real data

We also compared MCALIGN2 with AVID and CLUSTALW using real intronic DNA sequences from mouse and rat. Although we do not know the true alignments for real sequence data, we can still judge the alignment performances of different methods by examining the plausibility of the alignments (e.g. positions of gaps in the alignments and proportion of matched bases). Here we show three specific cases in which MCALIGN2

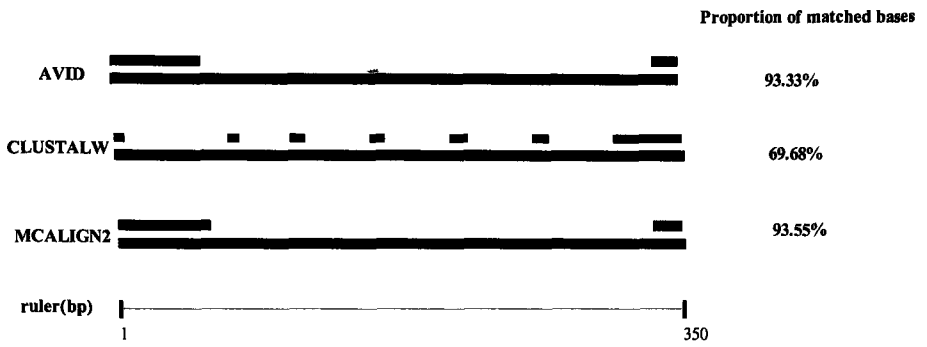
Table 2.5. Performance of MCALIGN2 and AVID compared by estimator of divergence time based on a General Time-Reversible Model.

Simulated		Proportion of matched bases				
t	Θ	Alignment Known	AVID	MCALIGN2(JC)	MCALIGN2(K2P)	MCALIGN2(GTR)
0.05	0.225	0.0504 (0.0100)	0.0523 (0.0181)	0.0501 (0.0100)	0.0500 (0.0100)	0.0500 (0.0100)
0.10	0.225	0.0991 (0.0145)	0.1023 (0.0159)	0.0981 (0.0151)	0.0976 (0.0149)	0.0983 (0.0149)
0.15	0.225	0.1522 (0.0191)	0.1575 (0.0226)	0.1496 (0.0192)	0.1494 (0.0192)	0.1502 (0.0194)
0.20	0.225	0.2034 (0.0225)	0.2131 (0.0283)	0.1985 (0.0238)	0.1978 (0.0239)	0.2003 (0.0235)
0.25	0.225	0.2531 (0.0286)	0.2699 (0.0383)	0.2493 (0.0316)	0.2453 (0.0308)	0.2491 (0.0311)
0.30	0.225	0.3003 (0.0302)	0.3222 (0.0420)	0.2944 (0.0324)	0.2914 (0.0311)	0.2987 (0.0323)
0.15	0.10	0.1520 (0.0207)	0.1528 (0.0219)	0.1510 (0.0203)	0.1508 (0.0203)	0.1510 (0.0204)
0.15	0.30	0.1509 (0.0200)	0.1634 (0.0301)	0.1489 (0.0207)	0.1470 (0.0208)	0.1499 (0.0207)
0.15	0.40	0.1526 (0.0208)	0.1794 (0.0398)	0.1507 (0.0213)	0.1477 (0.0203)	0.1508 (0.0211)

Estimates of sequence divergence, *t*, from 200 replicates for each combination of t and θ, with sequences of length 500 base pairs. Estimated root mean square error (e.r.m.s.e.) is shown after divergence time in parentheses. Here MCALIGN2 is tested by assuming either the correct model of DNA evolution (GTR) or the incorrect model (JC and K2P).

performed quite differently from AVID and CLUSTALW. The parameters for the rodent (mouse and rat) alignment model (θ and w) used in MCALIGN2 were described in Keightley and Gaffney (2003). They were estimated from 27 orthologous intron sequences of the closely related mouse species *Mus domesticus*, *Mus spretus*, and *Mus caroli*, for which nucleotide and indel divergences are sufficiently low as to make alignments by heuristic methods practically unambiguous (Keightley and Gaffney 2003).

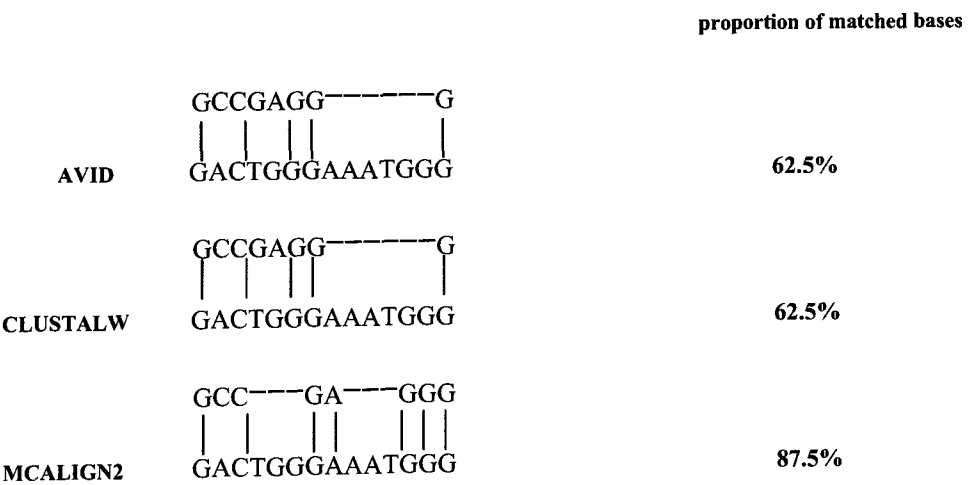
As shown in Figure 2.4(a), AVID and MCALIGN2 produced similar alignments, which include a long gap between ~70bp - ~320bp. However, the alignment produced by CLUSTALW has several small gaps, which are separated by small segments of align-


Proportion of matched bases
AVID
93.33%
CLUSTALW
69.68%
MCALIGN2
93.55%
ruler(bp)
1
350

(a)


Proportion of matched bases
AVID
88.19%
CLUSTALW
45.93%
MCALIGN2
91.86%
ruler(bp)
1
600
1200

(b)


proportion of matched bases
GCCGAGG-------G
AVID
GACTGGGAAATGGG
62.5%
GCCGAGG-------G
CLUSTALW
GACTGGGAAATGGG
62.5%
GCC---GA---GGG
MCALIGN2
GACTGGGAAATGGG
87.5%

(c)

Figure 2.4. Alignments of real non-coding DNA sequences from mouse and rat produced by AVID, CLUSTALW and MCALIGN2. The scale is shown below the alignments, and proportion of matched bases in the alignment is shown on the right. There are three cases (a) alignments of intronic DNA sequence from the *Fshb* gene, (b) alignments of intronic DNA sequences from the *Omd* gene and (c) alignments of small pieces of intronic DNA sequences from *Omd*, in which MCALIGN2 performed quite differently from the others.

ed bases. In this example, 93% of base pairs are matched in alignments produced by AVID and MCALIGN2, while only 70% of base pairs are matched in the alignment produced by CLUSTALW. Although it is impossible to say which alignment is the true alignment, the positions of gaps and proportion of matched bases can give someindications of the alignment plausibility. As the gap-open penalty is higher than the gap-extension penalty, the cost of having several small gaps is higher than the cost of having a long gap, if the total length of gaps is similar among different alignments. Meanwhile, as the match state has a positive effect on the alignment probability, the alignment with the higher proportion of matched bases is more likely to be correct. Therefore, from the point of view of the alignment probability, the alignments produced by MCALIGN2 and AVID in this case are more plausible.

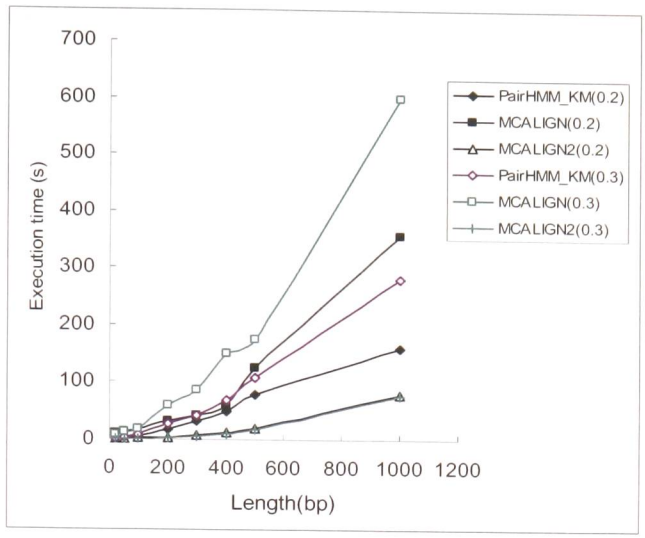
Figure 2.4(b) shows a different fragment from alignments produced by AVID, CLUSTALW and MCALIGN2. In this case, the alignment produced by MCALIGN2 also has a long gap from ~300bp - ~1000bp, and it has the highest proportion of matched bases compared to other alignments. However, the alignment produced by CLUSTALW has several small gaps and a long gap in the terminal portion, and it has the lowest proportion of matched bases. Although the alignment produced by AVID looks better

than the one produced by CLUSTALW, it is still fragmented by several small-length gaps.

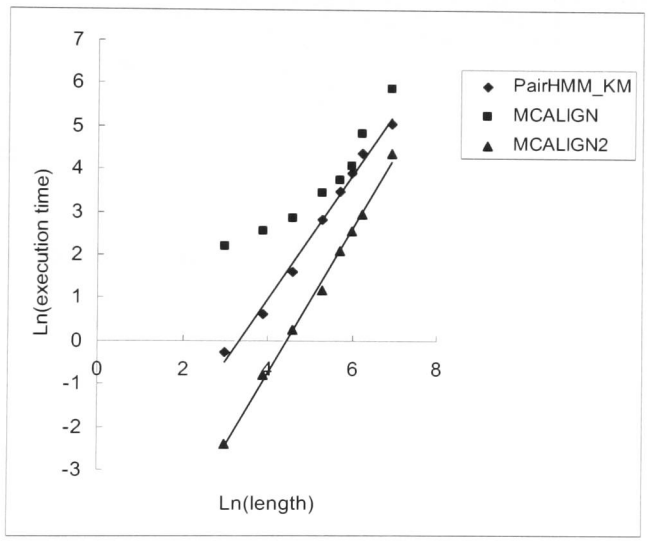
However, MCALIGN2 does not always produce fewer gaps than other methods. As shown in Figure 2.4(c), the alignment produced by MCALIGN2 has more gaps than the others, but a smaller number of nucleotide differences. Without other information it is impossible to say which is more plausible.

Execution time

Figure 2.5 shows the execution times of the different alignment algorithms tested above, as a function of sequence length. Execution times were measured in on a 2.8GHz Intel ® Xeon™ processor. Results are shown for divergence time t = 0.2 and 0.3, and ratio of indels to point substitutions θ = 0.225. Figure 2.5 shows that for the DP method (MCALIGN2), execution time increases as a quadratic function of sequence length, as expected. Similar behaviour is observed for the PairHMM_KM method, since this calculates the sum of the probabilities of all alignments for two given sequences using the forward algorithm for pair HMMs, and this gives a time and memory complexity on the order of L^2 (L be the length of the sequence) (Knudsen and Miyamoto 2003). For the MC method (MCALIGN) execution time increases with sequence length, and, although it does not follow a power law, it is roughly quadratic for long sequences. Although it shows a substantial improvement over the previous Monte Carlo method and it is faster than the PairHMM_KM method, MCALIGN2 still cannot compare with a heuristic alignment method such as AVID. To align two 1000-bp sequences (t = 0.2) takes about



(a)



(b)

Figure 2.5. Execution time plotted against sequence length for sequence divergence of 0.2 and 0.3, and ratio of indels of 0.225. Execution times were estimated from the average of ten simulations. (a) Execution time of MCALIGN2 comparing to MCALIGN and the pair HMM method of Knudsen and Miyamoto (black points and lines for divergence of 0.2; colored points and lines for divergence of 0.3). (b) Slope of log(execution time) against log(length) for sequence divergence of 0.2. All the numbers are in natural log. The slope of log(time) against log(length) is 1.91 for MCALIGN2, which means the program closely follows the expected

algorithm time operation $O(L^2)$. The slope of log(time) against log(length) for PairHMM_KM is 1.78, not far away from the order L^2. The pattern for MCALIGN is hard to track.

0.1s using AVID, about 80s using MCALIGN2, about 160s using PairHMM_KM and about 350s using MCALIGN. Furthermore, it is also shown in Figure 2.5(a) that both MCALIGN and PairHMM_KM take longer to align sequences with larger divergence times, whereas execution time of MCALIGN2 is unaffected by divergence. However, given the pair HMM model used in PairHMM_KM, execution time should not be affected by divergence for this method. We suppose that this occurred due to small tolerances chosen for ML estimation of divergence time in this program.

2.5. Discussion

The problem of statistical inference of an alignment can be separated into two parts: specifying a scoring function, and finding an alignment that optimises that scoring function. The scoring function is specified on biological and/or statistical grounds, and determines the biological meaningfulness and accuracy of the inferred alignment. The choice of optimising algorithm determines the speed of the method, and may hamper accuracy if convergence to a global optimum cannot be guaranteed. A useful alignment method must produce biologically meaningful and accurate alignments, and also must do so quickly. There is a trade-off because the most biologically realistic scoring functions are difficult to optimise.

Many scoring functions can essentially be described by the relative contributions for individual nucleotide substitution and indel events, which were traditionally thought of as penalty scores for mismatches and for gaps. However, no general theory guides the selection of these penalties (Reese and Pearson 2002), unless divergence time is known (Durbin et al. 1998). Although almost all scoring functions have a *probabilistic* interpretation (Durbin et al. 1998), only ones in which divergence time is an explicit parameter have an *evolutionary* interpretation. This inclusion of a time parameter is crucial in allowing us to train or parameterize our model using closely related sequences, in order to improve the accuracy of alignments between more distantly related sequences. Although the idea of training a scoring function on known alignments is an old one (especially with respect to amino acid substitutions (e.g. PAM250 matrix of Dayhoff et al. (1978))), in the past it has generally been necessary to use a training set of sequences at similar evolutionary distance as the sequences that are ultimately to be aligned.

Heuristic scoring functions are often chosen because an algorithm exists to optimize them efficiently. However, without any underlying evolutionary model, the alignments produced by such methods will be biased (at least at some evolutionary distances), in the sense that they will exhibit features that depart in a systematic direction from the true alignment.

The evolutionary model used in our method strikes a balance between biological realism and computational tractability. We ignore multiple hits of indel events, and assume a distribution of indel lengths that corresponds to an improved affine gap penalty scheme. Our model is therefore quite different from more realistic evolutionary models

that account properly for multiple hits of indel events (Thorne et al. 1991; Thorne et al. 1992; Knudsen and Miyamoto 2003; Miklos et al. 2004). The TKF91 model is particularly unrealistic for non-coding DNA, since it allows only single base indels. Keightley and Johnson (2004) suggest that the present model (ignoring multiple hits for indels) is a better approximation to their simulation model (which allowed multiple hits of multi-base pair indels), for the parameter values used in their simulations. The TKF92 model allows a geometric distribution of indel lengths, but only allows whole insertions to be subsequently deleted, or vice versa. That model has therefore been criticised as introducing non-biological "hidden fragment boundaries". Since our model does not allow insertions to be deleted at all, or vice versa, it could be seen as also introducing "hidden fragment boundaries". Our model allows a more realistic distribution of indel lengths than the TKF92 model. The approach of Knudsen and Miyamoto (2003) could be seen as an extension of the TKF92 model, assuming a geometric distribution of indel lengths and allowing multiple hits involving up to two indel events. Our results suggest that this model (approximated using a three state pair HMM), and our model (using a seven state HMM) offer approximations of very similar quality. Intuitively, we would have expected our model to be superior when multiple hits of indel events were rare, i.e. for relatively smaller evolutionary distances and indel rates. However, it seems that in such cases the performance of both methods is so good that it is hard to detect any difference. The "long indel" model of Miklos et al. (2004) is certainly more realistic than either model, since it allows an arbitrary distribution of indel lengths and accounts almost exactly for multiple hits of indels. However, the finite trajectory algorithm

(Miklos et al. 2004) used to account for multiple hits is computationally expensive ($O(L^4)$in complexity).

When comparing the present method (MCALIGN2) against a previous Monte Carlo approach (MCALIGN (Keightley and Johoson 2004)) we are comparing the performance of two different optimisers, with the same scoring function. Generally MCALIGN2 has better global optimum finding properties, and is much faster than the Monte Carlo method to align the same sequences. There are two major reasons for this improvement:

(1) MCALIGN2 uses a dynamic programming algorithm that is guaranteed to find the most probable alignment for a given divergence time, whereas the stochastic hillclimbing algorithm used in the Monte Carlo method can only search locally by making heuristically chosen adjustments to an alignment.

(2) MCALIGN2 stops its search when the maximising divergence time is bracketed to high precision, with the bracket length being reduced by a geometric factor at each step of the algorithm. In contrast, the Monte Carlo method must search until no improvement in alignment probability is found during a predetermined number of iterations.

In comparisons of MCALIGN2 against the pair HMM method of Knudsen and Miyamoto, a method with an evolutionary time parameter and an affine gap penalty (Knudsen and Miyamoto 2003), we found that the two methods performed very similarly for almost all cases, but MCALIGN2 is computationally faster. When

comparing MCALIGN2 against AVID, a time-naive model (Bray et al. 2003), we found that MCALIGN2 produced better quality alignments than AVID for almost all combinations of parameters. This shows that, when the evolutionary model is known, this knowledge can be used in in a model based inference method to estimate alignment more accurately.

Despite being substantially faster than our original Monte Carlo approach and the pair HMM method of Knudsen and Miyamoto, MCALIGN2 cannot compete with AVID in terms of execution time, because of the clever heuristics used by AVID. Its general strategy for aligning two sequences is to select anchors using a variant of the Smith-Waterman algorithm (Smith and Waterman 1981) to split long sequences into short sequences, which are aligned by a dynamic programming algorithm, Needleman-Wunsch (Needleman and Wunsch 1970). A set of maximal matches between sequences is constructed using a suffix tree. This approach is fast and memory efficient, and practical for sequence alignments of large genomic regions up to megabases long (Bray et al. 2003). In principle, the fast heuristics used by AVID can be applied for any pair HMM, and therefore could be combined with our approach to give faster, high quality alignments.

In order to examine the robustness of the MCALIGN2 method, we also investigated cases in which the model assumed in the MCALIGN2 analysis was a simpler model (JC or K2P) than the model the data were simulated under (GTR). Generally, the MCALIGN2 method assuming an incorrect model still has good performance for small and medium divergence times, but for larger divergence times and/or higher indel rates,

performance suffers slightly compared with when the correct GTR model was assumed. Therefore, when aligning sequences from distant species, it is desirable to use an evolutionary model that is as realistic as possible. However, it is in precisely this situation that it may be most difficult to estimate a model, because the assumption that the evolutionary process is the same between closely and distantly related species is most likely to break down.

When inferring alignment in a Bayesian framework, divergence time is a nuisance parameter that must be eliminated by integration (Equation 2.2). The computational implementation of our method relies totally on being able to approximate this integral (Equations 2.4 and 2.6) rather than having to calculate it numerically using e.g. quadrature. The approximations we make will be good when $P(t|a,S)$ is approximately normal with constant variance for a certain set of high probability alignments. Because $P(t|a,S)$ is a product of multinomial probabilities, the normality approximation will be good for long sequences under most models of molecular evolution. The assumption of constant variance will be reasonable when high probability alignments differ from each other by only a few indels and substitutions, relative to the total sequence length. As a concrete check of this assumption, we used the Monte Carlo search algorithm of Keightley and Johnson (2004) and retained the set of all alignments visited that had probability at least 0.01 as large as the maximum probability. Within this set, the correlation between $P(a|S)$ computed "exactly" (using quadrature) and $P(a|S)$ exceeded 0.98.

It is worth mentioning that, to our knowledge, no better method has been found for eliminating divergence time as a nuisance parameter when estimating alignment. Most authors concentrate on finding the true MLE for t, summing over all possible alignments, using the EM algorithm (Thorne et al. 1991; Thorne et al. 1992; Miklos and Toroczkai 2001; Holmes and Bruno 2001). The best way to estimate the alignment has not been considered in detail, but a common approach is to use the most probable alignment conditional on the observed sequences and conditional on the MLE for t. Although our method has a more direct Bayesian justification, given the approximations made it is likely that the two approaches will give similar results.

2.6. Conclusions

Sequence alignment is a major issue for the evolutionary analysis of non-coding DNA. We developed a model-based method, MCALIGN2, as an improvement to the previous Monte Carlo method MCALIGN. MCALIGN2 uses a deterministic global optimiser to find the alignment with the highest posterior probability. It allows a rich class of evolutionary models of indel length along with any time reversible model of nucleotide substitution. As shown in the test results, MCALIGN2 outperforms other available non-coding DNA sequence alignment methods for all the cases we have considered.

2.7. Availability and requirements

Project name: MCALIGN2

Project home page: http://homepages.ed.ac.uk/eang33/

Operating system: Platform independent

Programming language: C++

Other requirements: C++ compiler if downloading and compiling the source code

Licence: FSF GENERAL public licence.

2.8. Acknowledgements

We thank Dr Bjarne Knudsen for providing their pair HMM program, Daniel Halligan for some useful comments and Daniel Gaffney for providing mouse and rat intronic sequences. JW was supported by Dorothy Hodgkin Postgraduate Studentship Award. TJ was supported by the Biotechnology and Biological Sciences Research Council grant #206/D16977.

Chapter 3.

Effect of divergence time and recombination rate on molecular evolution of *Drosophila* INE-1 transposable elements and other candidates for neutrally evolving sites

Jun Wang, Peter D. Keightley and Daniel L. Halligan*

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK.

***Corresponding author:** Daniel L. Halligan, e-mail: Daniel.Halligan@ed.ac.uk

JW and DLH initially conceived of and designed this study. JW performed all the data analysis, except the extraction and analysis of the FEI, all four-fold sites, which were carried out by DLH in a separate project. PDK co-designed and commented on the study. JW wrote the manuscript. PDK and DLH commented on and wrote the manuscript. All authors read and approved the final manuscript.

3.1. Abstract

Interspecies divergence of orthologous transposable element remnants is often assumed to be simply due to genetic drift of neutral mutations that occurred after the divergence of the species. However, divergence may also be affected by other factors, such as ancestral polymorphisms, or selection. Moreover, variation in mutation rate may also contribute to the difference in the divergence among elements from different regions. Here, we attempt to determine the impact of these forces on divergence of three classes of sites that are often assumed to be selectively unconstrained (INE-1 TE remnants, sites within short introns and four-fold degenerate sites) in two different pairwise comparisons of *Drosophila* (*D. melanogaster vs. D. simulans* and *D. simulans vs. D. sechellia*). We find that divergence of these three classes of sites is strongly influenced by the recombination environment in which they are located, and this is especially true for the closer *D. simulans vs. D. sechellia* comparison. We suggest that this is mainly a result of the contribution of ancestral polymorphisms in different recombination regions. We also find that intergenic INE-1 elements are significantly more diverged than intronic INE-1 in both pairwise comparisons, implying the presence of either negative selection or lower mutation rates in introns. Furthermore, we show that substitution rates in INE-1 elements are not associated with the length of the noncoding sequence in which they are located, suggesting that reduced divergence in long noncoding sequences is not due to reduced mutation rates in these regions. Finally, we show that GC content for each site within INE-1 sequences has evolved towards an equilibrium value (~33%) since insertion.

3.2. Introduction

Transposable elements (TEs) are mobile, repetitive DNA sequences that are major components of many host genomes. TEs can be divided in two major classes, Class I (RNA mediated) and Class II (DNA mediated) (Berg and Howe 1989; McDonald 1993). Class I elements, or retrotransposons, are transcribed into RNA, and then reverse transcribed and reintegrated into the genome, thereby duplicating the elements. Class II elements, or DNA transposons, are generally excised from one genomic site and integrated at another by a "cut and paste" mechanism involving a transposase (see review in Kazazian 2004). TEs comprise a large part of many eukaryotic genomes including mammals, such as human (~50%) and mouse (~40%) (Venter et al. 2001; Deininger and Batzer 2002; Waterston et al. 2002). In contrast, it is estimated that only 5.3% of the *Drosophila melanogaster* euchromatic genome is comprised of TE insertions (Quesneville et al. 2005).

The most abundant family of TEs in the *D. melanogaster* subgroup is INE-1 (Kapitonov and Jurka 1999, 2003). Kapitonov and Jurka (2003) speculated that INE-1 is a remnant of the *D. virilis Penelope* retrotransposon, but upon closer inspection it appears that INE-1 is more likely to be a family of nonautonomous DNA transposons (Pyatkov et al. 2002; Slawson et al. 2006; Yang et al. 2006). Previous analyses of this element have indicated that a burst of INE-1 transposition occurred in the ancestor of the *D. melanogaster* species complex ~5-10 Myrs ago, and the element is thought to have remained inactive ever since, at least in the *D. melanogaster* lineage (Kapitonov and

Jurka 1999; Kapitonov and Jurka 2003; Singh and Petrov 2004; Singh et al. 2005a). Thus, it is reasonable to assume that many of these transposable element remnants are selectively unconstrained, and their evolution has therefore been used to assess substitution rate heterogeneity (*i.e.*, background substitutional patterns) in the *D. melanogaster* genome (Singh et al. 2005a).

In the *Drosophila* genome, there are several other classes of nucleotide sites that have been hypothesized to be evolving close to neutrally. One class are sites in short introns (<65bp in total length), outside splice control regions (base pairs 8–30 from the 5′ end), which have been termed the fastest evolving intronic (FEI) sites (Halligan and Keightley 2006). A second are four-fold degenerate sites within coding sequences. However, there is evidence for weak selection for translational efficiency on at least some four-fold degenerate sites, resulting in codon-usage bias in *Drosophila* (Akashi 1995; McVean and Vieira 2001). Although sites within INE-1 elements and the two other classes of sites have all been assumed to be evolving close to neutrally in the past, there has been no attempt to compare their patterns of evolution.

Molecular evolution is expected to differ between these classes of sites for several reasons, and, even if sites are completely selectively unconstrained, there are several forces that could affect their interspecies divergence. Firstly, divergence between a pair of species is not only affected by the time since speciation but also by the time to coalescence (which is subject to stochastic variance) of those differences that are the result of polymorphisms in the common ancestor. Since the frequency of recombination affects the effective population size (N_e) of a region of the genome via Hill-Robertson

interference (Hill and Robertson 1966; Felsenstein 1974) or via the action of selective sweeps (Maynard-Smith and Haigh 1974; Kaplan et al. 1989), regions of low recombination are expected to have lower N_e compared to regions of high recombination. There is evidence to suggest that this is indeed the case in *Drosophila* (*e.g.* Betancourt and Presgraves 2002, Presgraves 2005, Haddrill et al. 2007). This will lead to an accelerated rate of genetic drift in regions of low recombination, and thus the coalescent time for polymorphic sites in the common ancestor will tend to be shorter, leading to lower divergence. This effect is expected to be strongest when the ratio of polymorphism (in the ancestor) to fixed differences (between species) is high. Secondly, interspecies divergence of unconstrained sequences can be affected by variation in the mutation rate. For example, it has been suggested that recombination itself may be mutagenic (Lercher and Hurst 2003; Hellmann et al. 2003; Yi et al. 2004). Finally, the mutation rate may be affected by transcription-coupled repair. In this case, the mutation rate and therefore divergence would be expected to be lower in sections of the genome that are transcribed.

If sites are under weak selection, then interspecies divergence can also be affected by variation in the recombination rate, since selection acting on these sites is expected to be less effective in regions of the genome with lower effective population size (Hill and Robertson 1966; Kliman and Hey 1993; Hey and Kliman 2002). This will lead to increased divergence in low recombination rate regions, if there is weak negative selection, but could produce the opposite pattern if there is a substantial fraction of positively selected substitutions. It is often assumed that TE remnants are unconstrained,

there is mounting evidence that selection acts on at least some TEs. For example, several recent studies have indicated that some TEs may cause changes in gene regulation in plants (White et al. 1994) and mammals (McDonald 1993; Deininger et al. 2003; Jordan et al. 2003; van de Lagemaat et al. 2003). In *Drosophila*, two classes of non-LTR retrotransposons of *Drosophila melanogaster*, Het-A and TART, integrate at specific ribosomal RNA gene locations, and maintain the telomeres of *D. melanogaster* chromosomes (Jakubczak et al. 1990; Pardue and DeBaryshe 2003). Also, the expression of the 17.6 retrotransposon in *Drosophila* has been shown to be crucial for the development of some tissues, including the eyes (Mozer and Benzer 1994). However, compared to mammals, there are relatively fewer fixed TE insertions in *Drosophila*, and segregating TE insertions appears to be at very low frequencies (Charlesworth and Langley 1989), suggesting that most TE insertions are deleterious. Thus, in contrast to mammals, it is possible that a relatively high proportion of fixed *Drosophila* TE insertions may be adaptive and it is interesting to test whether TE remnants, particularly those that persist for long periods of evolutionary time, are functional or not.

Furthermore, examining the molecular evolution of INE-1 elements could also shed light on the forces operating on unique noncoding DNA. For example, Haddrill et al. (2005) and Halligan and Keightley (2006) have claimed that there is a substantial amount of negative selection operating on noncoding DNA in *Drosophila*, based on reduced divergence in long intronic and intergenic sequences compared to synonymous sites and short introns. This raises the question of whether this pattern can instead be

explained by processes other than selection, *e.g.* lower mutation rates in long noncoding sequences. If so, a similar reduction in divergence would be expected to be observed in INE-1 elements located within long noncoding sequences.

In this paper, we investigate the molecular evolution of INE-1, and compare it to that of FEI and four-fold degenerate sites, by calculating mean divergence between *D. melanogaster* and *D. simulans* as well as between *D. simulans* and *D. sechellia* for sections of the genome that have different frequencies of crossing over. We look for evidence of variation in the substitution rate by testing for over-dispersion of substitutions in INE-1 elements within each recombination category. We investigate differences between intergenic and intronic INE-1 elements. We also investigate the relationship between divergence of INE-1 elements and the lengths of the noncoding sequences in which they reside, to test whether the difference in rates of substitution observed between noncoding sequences of different lengths in *Drosophila* can also be observed within the INE-1 elements located in these sequences. Finally, we investigate patterns of base composition and point substitution within extant copies of INE-1 in *D. melanogaster* and *D. simulans,* by comparing them to inferred ancestral sequence and calculate the equilibrium GC content in the *Drosophila* genome.

3.3. Materials and Methods

Compilation and alignment of sequences

We generated three data sets of pairwise INE-1 alignments. The first (data set 1) was derived from *D. melanogaster* and *D. simulans* noncoding sequence alignments of Halligan and Keightley (2006). These noncoding alignments cover ~80 Mb of genome sequence and were initially aligned using MAVID (Bray and Pachter 2004), then refined by realigning with MCALIGN2 (Wang et al. 2006). The second data set (data set 2) was also comprised of *D. melanogaster* and *D. simulans* pairwise alignments, but was extracted from independently collected three-way alignments of *D. melanogaster*, *D. simulans* and *D. sechellia*. The final set (data set 3) was comprised of *D. simulans* and *D. sechellia* pairwise alignments obtained from the same three-way alignments used for data set 2. The divergence between *D. simulans* and *D. sechellia* is much lower than that between *D. melanogaster* and *D. simulans*, allowing us to make inferences about processes whose magnitude of effect depends on divergence time. Data set 2 is a subset of data set 1, which provides a control for data set 3, since they are obtained from the same three-way alignments and are comprised of the same orthologous INE-1 elements.

We used a similar method to that described by Halligan and Keightley (2006) to obtain the three-way noncoding alignments used to derive data sets 2 and 3. We obtained a list of all currently annotated *D. melanogaster* genes from NCBI's Entrez Gene (Release 4.1) (excluding RNA genes and poorly annotated genes). We then used reciprocal best-hits BLAST to identify the locations of orthologous exons in the *D. simulans* (the April 2005 consensus assembly from the Genome Sequencing Center WUSTL School of Medicine) and *D. sechellia* (the October 2005 assembly from by the Broad Institute of

MIT and Harvard) genome sequences. We attempted to remove any genes from the data set for which the exon/intron structure may have changed. In order to do this we chose to only use genes where we could recover all exons by the reciprocal best-hits BLAST method, located in the same order as those in *D. melanogaster*. This should ensure that the exon/intron order in our selected genes is the same between species. Furthermore, we removed genes from the data set if the coding sequences (CDS) were invalid (a CDS was considered valid if it started/ended with a start/stop codon, had no internal stop codons and was a multiple of 3bp in length). The start and end positions of the located exons were then used to extract the adjacent intronic and intergenic sequences. Noncoding DNA sequences were only extracted if a reciprocal-best hit for the two flanking exons was found in both *D. simulans* and *D. sechellia*. As a result of this stringency, these noncoding sequences cover only ~40Mb of the *D. melanogaster* genome sequence. Finally, we aligned orthologous sections of noncoding DNA using MAVID (Bray and Pachter 2004). We removed introns from the data set if they did not start or end with a 2bp consensus sequence and/or if the intron sequences in *D. simulans*/*D. sechellia* started or ended with gaps (these represent incorrectly aligned sequences under the assumption that the 2bp consensus should be aligned).

For each of the genomic noncoding sequence alignments (pairwise and three-way), we extracted the alignments of FEI sites (base pairs 8–30 from the 5′ end of introns <65bp in length, see Halligan and Keightley 2006). We also aligned the identified orthologous exons between *D. melanogaster* and *D. simulans* and orthologous exons in all three

Drosophila species using the amino-acid alignment obtained from CLUSTALW (Thompson et al. 1994). We then extracted the alignments of four-fold degenerate sites.

Identification and extraction of orthologous INE-1 elements

We extracted orthologous INE-1 elements from the noncoding pairwise and three-way noncoding alignments as follows:

(1) We removed gaps from the alignments and used the reported consensus sequence for INE-1 (Kapitonov and Jurka 2003) in RepeatMasker (http://www.repeatmasker.org) to identify INE-1 elements in all species.

(2) We found the locations of the INE-1 elements identified with RepeatMasker within the noncoding alignments, and extracted only the sections of the alignments identified as INE-1 in all species.

(3) We excluded all alignments with fewer than 50 valid bases (*i.e.*, A, T, G or C) in any species or 100 alignment columns.

(4) We attempted to exclude non-homologous sections within the INE-1 alignments by masking sections in which there were short lengths of bases surrounded by long gaps or in which divergence was above 0.30 (between *D. melanogaster* and *D. simulans*) or above 0.12 (approximately three times the mean interspecies divergence between *D. simulans* and *D. sechellia*) within a 50bp sliding window. We also excluded alignments altogether if these masked sections comprised more than 60% of the alignment.

For data set 1, we identified 1,657 and 1,581 INE-1 elements in step 1 from *D. melanogaster* and *D. simulans*, respectively, of which 1,103 remained after step 2. This was reduced to 613 and 353 after steps 3 and 4, respectively (312 from intergenic regions and 41 from intronic regions). For data sets 2 and 3, step 1 identified 400, 395 and 381 INE-1 elements from *D. melanogaster*, *D. simulans*, and *D. sechellia*, respectively. This was reduced to 161 after step 2. After steps 3 and 4, data set 2 comprised of 149 *D. melanogaster* and *D. simulans* INE-1 alignments (91 intergenic and 58 intronic) and data set 3 comprised of 161 *D. simulans* and *D. sechellia* alignments (99 intergenic and 62 intronic). Details of the results from this extraction procedure are available by request.

Recombination regions

Each data set of INE-1 elements, FEI, and four-fold degenerate sites was sorted according to cytological map location, and divided into categories with high, intermediate and low frequencies of crossing over, and a group with no crossing over, based on the regions described in Charlesworth (1996) and listed in Haddrill et al. (2007). These cytological locations are based on the band coding system of Charlesworth and Lapid (1989) and Charlesworth et al. (1992), which assigns approximate physical positions to loci for which information on DNA variability is available. We removed some bands in telomeric and centromeric polytene regions due to a lack of experimental recombination data. We attempted to apply estimates of recombination rate for *D. melanogaster* to its sister species, *D. simulans* and *D. sechellia* with some modification. We have attempted to limit the effects of changes in the

recombination environment by excluding sequences from a number of cytological bands due to uncertainty over changes in the recombinational environment between *D. melanogaster* and *D. yakuba* (Marais et al. 2004; Haddrill et al. 2007). Furthermore, we can be fairly confident that chromosome 4 has remained non-recombining since the split of the *D. melanogaster* species complex (Jensen et al. 2002).

To test the association between crossing-over frequency and divergence, we assigned values 4, 3, 2 and 1 to the crossing-over frequency classes, high, intermediate, low and no, respectively, and calculated the Spearman rank correlation between divergence and crossing-over frequency value.

Estimating divergence and equilibrium GC content of INE-1 elements

All divergence estimates were corrected for multiple hits (Kimura 1980). Mean divergence for pairwise comparisons between species was calculated on a per site basis, and 95% confidence intervals were calculated by bootstrapping by TE element. To compare mean divergence and GC content between classes of sites, we generated 1,000 bootstrap estimates of the statistic for each of the classes to be compared and then calculated the difference between each of the 1,000 bootstrap estimates from the two classes and tested whether the distribution of these differences was significantly different from zero.

We inferred the polarity of substitutions along the lineages leading to the extant INE-1 sequences using the consensus sequence as an outgroup and used this information to estimate the expected proportion of bases that are G/C at equilibrium (p_{GC}). Let the

substitution rates from G:C to A:T pairs and A:T to G:C pairs be r_{GC} and r_{AT}, respectively. Under the assumption that the pattern of substitution rates is stable through time, at equilibrium, the number of G:C bases replaced by A:T bases will be equal to the number of A:T bases replaced by G:C pairs, so $r_{GC}p_{GC} = r_{AT}(1 - p_{GC})$. Therefore, the GC content at equilibrium $p_{GC} = r_{AT}/(r_{GC} + r_{AT})$. For each recombinational category, we randomly sampled INE-1 elements with size equal to the number of elements in the sampled category with replacement. For each sampled element, we observed the number for each type of single nucleotide substitution. We then calculated the total rate of single nucleotide substitutions and the expected GC content at equilibrium for these sampled elements. We conducted this process 1,000 times, and obtained the mean and 95% confidence intervals for rates of different nucleotide substitutions and the expected GC content at equilibrium.

Measurement of dispersion of substitutions among INE-1 elements

If substitutions result from independent mutations occurring along each lineage since time of speciation without any effect of selection, then each INE-1 element is expected to accumulate substitutions at the same rate. The observed number of nucleotide differences in each element is then expected to follow the same binomial distribution (*i.e.* the probability of observing a difference at a site will be constant across elements). We tested this null hypothesis by assuming that the probability of observing a difference at any site was equal to the observed mean divergence (uncorrected for multiple hits). For each recombination category in each data set, we generated 1,000 simulated data sets consisting of divergence estimates for sequences with the same length distribution

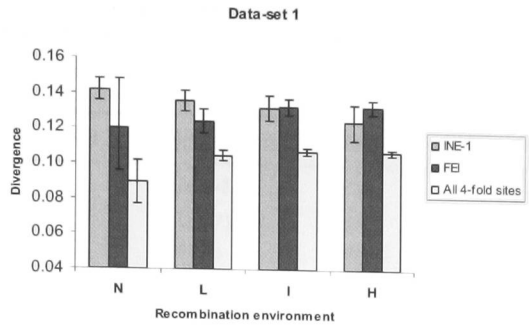
as the actual data and calculated the variance of divergence between elements for each simulation to generate a null distribution of variance of divergence amongst elements. If substitutions are over-dispersed, then the observed variance for the given data set will be larger than that expected under the null hypothesis of equal rates.

3.4. Results

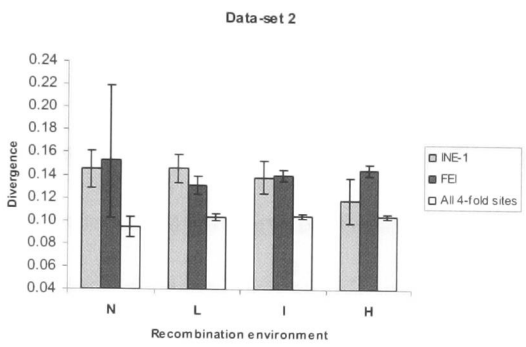
Evaluation of mean divergence controlled for recombination rate

Previous analysis of INE-1 elements in *D. melanogaster* (Singh et al. 2005a) has suggested that there are differences in rates of substitutions in INE-1 copies amongst different recombination environments. To investigate this further, we divided INE-1, FEI and four-fold degenerate site alignments into four crossing over frequency categories (high, intermediate, low and no crossing over) and compared divergence of each site class within and amongst categories. We found that the divergence of four-fold degenerate sites is always lower than that of FEI sites (see Figure 3.1). This confirms previous observations (Halligan and Keightley 2006) and is consistent with reports that some four-fold degenerate sites are under weak selective constraints in *Drosophila* (Akashi 1995; McVean and Vieira 2001).

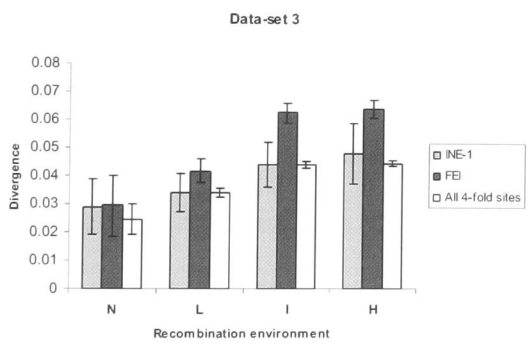
Interestingly, Spearman rank correlations between the divergence of FEI/four-fold sites and crossing over frequency are positive in all three data sets (Table 3.1). These correlations are significant for the cases of all sequence categories for data set 3 and FEI sites in data set 2. There are several possible explanations for the positive correlation



(A)



(B)



(C)

Figure 3.1. - Mean divergence (+/– 95% confidence intervals) of INE-1, FEI and four-fold sites from A, data set 1; B, data set 2; C, data set 3. Within each data set, the three classes of nucleotide sites were sub-divided into categories according to frequency of crossing over: high (H), intermediate (I), low (L) and no (N) crossing over. 95% confidence intervals were obtained by bootstrapping by element/intron/gene.

Table 3.1. Spearman correlation between divergence and rate of recombination for INE-1, FEI and four-fold degenerate sites. Here, we assigned 4, 3, 2 and 1 to be values for high, intermediate, low frequency of crossing over and no crossing over. The sample size is shown in the

Data Set	Spearman Correlation Coefficient (r)		
	INE-1	**FEI**	**Four-fold**
1	–0.171 **	0.032	0.015
	(353)	(3411)	(5130)
2	–0.153	0.054 **	0.003
	(149)	(3226)	(4955)
3	0.196 *	0.107 ***	0.12 **
	(161)	(3226)	(4955)

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

between divergence and crossing over frequency, such as stronger positive selection in high recombination regions, mutagenic effect of recombination, and differences in the level of the ancestral polymorphism (see Discussion for details).

However, divergence patterns of INE-1 elements are quite different (Table 3.1). In data sets 1 and 2, there is a negative correlation between divergence and crossing over frequency for INE-1 elements (significantly so for the larger data set 1), but a significantly positive correlation for INE-1 elements in data set 3. Furthermore, in regions of high recombination, INE-1 elements have a lower mean divergence than FEI sites (in all three data sets), and this is significant for data sets 2 ($p = 0.042$) and 3 ($p = 0.01$). It is possible to explain this latter result without the need to invoke selection on either FEI sites or INE-1 elements, if INE-1 elements were less polymorphic than FEI sites in the common ancestor.

In data set 3, consistent with the results for FEI and four-fold degenerate sites, there is a positive correlation between divergence and crossing over frequency for INE-1. This result could be explained by the fact these species are so closely related that the effect of increased coalescence time in the common ancestor in regions of high crossing over frequency dominates any other effects that could potentially cause a negative correlation.

It is worth mentioning that GC content and recombination rate are negatively correlated for the X chromosome, while the opposite is true for the autosomes (Singh et al. 2005b). It is therefore possible that our estimates of the correlation between divergence and crossing-over frequency are influenced by grouping X-linked elements and autosomal elements together. To investigate this, we divided the 353 INE-1 elements from data set 1 (since it is the largest) into subsets of X-linked elements (46) and autosomal elements (307). We then repeated the analysis on the autosomal elements only, and obtained very similar results to those for the grouped autosomal and X-linked elements. Unfortunately, there were insufficient X chromosome data to obtain reliable estimates of the correlation between divergence and crossing-over frequency, but we did not find any significant difference in mean divergence and GC content between X-linked and autosomal elements (data not shown).

Over-dispersion of INE-1 substitutions

Variation in the substitution rate can be assessed by testing for over-dispersion of substitutions. This variation could be caused by mutation rate variation, positive or negative selection that varies between elements or the contribution of polymorphism in

Table 3.2. Test of over-dispersion of substitutions in INE-1 elements for different crossing over categories in data sets 1, 2 and 3. Within each data set, INE-1 elements were subdivided according to different recombination environments. We show the null distribution of variance in divergence uncorrected for multiple hits with a constant substitution rate and the observed variance.

Data Set	Category	N	Mean Null Variance [95% Range]	Observed Variance
1	H	52	0.0007 [0.0004 – 0.0010]	0.0010 **
	I	100	0.0009 [0.0004 – 0.0009]	0.0009 *
	L	127	0.0007 [0.0005 – 0.0009]	0.0011 ***
	N	66	0.0007 [0.0004 – 0.0010]	0.0007
2	H	27	0.0015 [0.0006 – 0.0025]	0.0027 *
	I	43	0.0015 [0.0007 – 0.0023]	0.0017
	L	60	0.0014 [0.0008 – 0.0020]	0.0021 *
	N	16	0.0012 [0.0001 – 0.0023]	0.0008
3	H	28	0.0006 [0.0002 – 0.0009]	0.0011 **
	I	47	0.0005 [0.0003 – 0.0008]	0.0019 ***
	L	67	0.0003 [0.0002 – 0.0005]	0.0006 ***
	N	16	0.0003 [0.0001 – 0.0005]	0.0013 ***

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

the ancestor to divergence between species. The latter effect is expected to be strongest when the ratio of ancestral polymorphism to divergence is highest. We tested for over-dispersion in each crossing over category of each data set by comparing the observed variance in divergence across elements to the expected, assuming a constant substitution rate and the same alignment length distribution. We found evidence for over-dispersion of the substitution rate in 9 out of 12 cases (Table 3.2), the strongest evidence for over-

dispersion coming from elements in data set 3, on the basis that the ratio of observed to expected variance was highest and the p-values were the most significant. Although all of the above-mentioned processes could contribute to the patterns observed, only the third explanation, *i.e.* a contribution from ancestral polymorphism, clearly predicts that over-dispersion would be strongest in the closer pairwise comparison of *D. simulans* and *D. sechellia*.

Comparisons between intergenic and intronic INE-1 elements

One possible factor that could explain some over-dispersion of substitutions, either because of differences in the mutation rate, or differences in the level of selection, is transcription. A recent study in mammals found that transposable elements in introns have a slightly lower mean divergence than those in intergenic DNA (Gaffney and

Table 3.3. Mean divergence and GC content of intergenic and intronic INE-1 elements for data sets 1, 2 and 3. 95% confidence intervals for divergence and GC content were calculated by bootstrapping 1,000 times by INE-1 element.

Data Set	Class	N	Mean Divergence [95% CI]	GC Content [95% CI]
1	Intergenic	312	0.135 [0.131 – 0.139]	0.376 [0.372 – 0.380]
	Intronic	41	0.123 [0.122 – 0.135]	0.371 [0.361 – 0.381]
2	Intergenic	91	0.153 [0.142 – 0.163]	0.378 [0.368 – 0.388]
	Intronic	58	0.123 [0.114 – 0.132]	0.385 [0.372 – 0.398]
3	Intergenic	99	0.039 [0.034 – 0.045]	0.375 [0.363 – 0.388]
	Intronic	62	0.033 [0.028 – 0.039]	0.391 [0.378 – 0.404]

Keightley 2006). We tested for differences in substitution rate between intergenic and intronic INE-1 alignments in each data set. Consistent with the observations in mammals, intergenic INE-1 elements have somewhat higher mean divergence than intronic elements, and this is significant for data sets 1 ($p < 0.05$) and 2 ($p < 0.01$) but not 3 ($p = 0.28$) (Table 3.3). Furthermore, the difference between intergenic and intronic elements is consistent within each recombination category for all data sets (data not shown), suggesting that this result is not due to differences in levels of crossing over between intronic and intergenic regions. Additionally, there are no significant differences in mean GC content or mean length between intergenic and intronic elements in all three data sets.

Correlation between INE-1 divergence and length of noncoding sequence

It has previously been shown that there is a strong negative correlation between intron length and divergence in *Drosophila* (Haddrill et al. 2005; Halligan and Keightley 2006), and a similar pattern has also been found in intergenic sequences (Halligan and Keightley 2006). If these length correlations are the result of lower mutation rate in long noncoding sequences, or other general factors affecting all sites within long noncoding sequences, we would predict that INE-1 elements in long noncoding sequences would have lower divergence than those in short noncoding sequences. However, we found no significant correlation between INE-1 element divergence and length of the noncoding sequence in which the elements are located in any of the data sets (Spearman correlation $r = 0.061$, $p = 0.25$ for data set 1; $r = 0.072$, $p = 0.18$ for data set 2; $r = 0.051$, $p = 0.28$ for data set 3). Note that there is a difference in sequence lengths between intronic and

intergenic sequences, so we separated intronic sequences from intergenic sequences, and calculated the correlation of INE-1 divergence and non-coding sequence length separately. However, we did not found any relationship between these two factors for data sets 1 (intergenic: Spearman correlation r = -0.001, p = 0.98; intronic: Spearman correlation r = 0.25, p = 0.12), 2 (intergenic: Spearman correlation r = 0.01, p = 0.88; intronic: Spearman correlation r = 0.26, p = 0.16) and 3 (intergenic: Spearman correlation r = - 0.02, p =0.64; intronic: Spearman correlation r = 0.18, p = 0.34). It is worth mentioning that correlation is always close to zero for intergenic sequences, but positive and quite strong for intronic sequences (although not significantly). Our results suggest that the difference in divergence between short and long noncoding sequences is not a result of any factors, such as mutation rate, that have general differential effects on long *vs.* short noncoding sequences, and instead supports the conclusion that the difference is due to stronger or more extensive negative selection in long noncoding sequences.

Evolution of GC Content of INE-1

If the ancestral INE-1 sequence were available we would be able to polarize substitutions in the *D. melanogaster* and *D. simulans* lineages. Although previous studies have assumed the reported consensus is a fair approximation to the ancestral sequence (Singh and Petrov 2004; Singh et al. 2005a), none have tested this assumption. Using data set 1, we reconstructed the consensus sequences of INE-1 from the extant copies in *D. melanogaster* and *D. simulans* separately. The divergence between these two reconstructed consensus sequences is very low (~0.005). None of the four pairwise

Table 3.4. Mean divergence between extant INE-1 copies in *D. melanogaster* and *D. simulans* and the consensus sequences constructed from each species.

Consensus Species	Extant Copy Species	Mean Divergence [95% CI]
D. melanogaster	*D. melanogaster*	0.1879 (0.1831 – 0.1928)
D. melanogaster	*D. simulans*	0.1841 (0.1807 – 0.1875)
D. simulans	*D. melanogaster*	0.1878 (0.1830 – 0.1927)
D. simulans	*D. simulans*	0.1839 (0.1804 – 0.1873)

comparisons between the extant sequences of either species and the consensus sequences constructed from either species are significantly different from one another (Table 3.4). This implies that the consensus sequence of INE-1 is a good approximation to the ancestral sequence.

Given the ancestral sequence and the extant copies of INE-1, we were able to investigate how GC content has changed in each species. We calculated GC content in the consensus and mean GC content in extant sequences in *D. melanogaster* within 50bp sliding windows across each INE-1 element. Figure 3.2 shows these results for high and no crossing over categories for elements in *D. melanogaster* from data set 1 (plots for other categories showed very similar patterns, see Supplementary Figure 3.1). Although GC content appears to have reduced in the *D. melanogaster* and *D. simulans* lineages since their common ancestor, there is substantial variation across the INE-1 element. GC content in the consensus varies between 10% and 60%, whereas mean GC content of the extant copies shows much less variation. By polarizing substitutions in the *Drosophila melanogaster* and *D. simulans* lineages using parsimony with the consensus sequence as

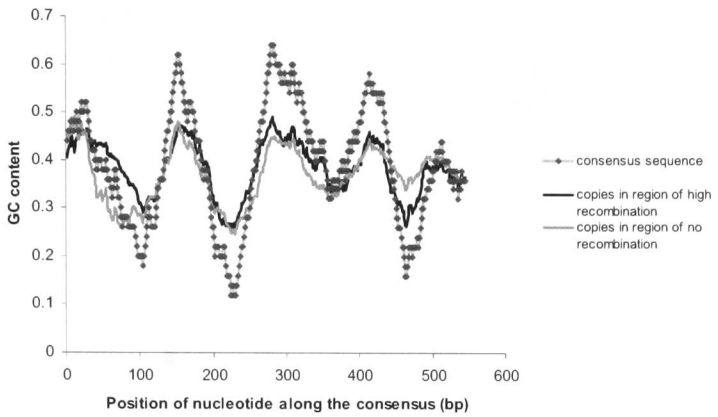


Figure 3.2. - Mean GC content of consensus sequence (light gray line with dark dots) and the orthologous INE-1 extant remnants in *Drosophila* for each nucleotide site along the consensus sequence in 50bp sliding windows in data set 1 for *D. melanogaster*. Within the data set, we only show results for INE-1 in regions of high (black line) and no (gray line) recombination.

an outgroup, we calculated lineage-specific rates of substitution. We estimated rates for each of 6 types of nucleotide substitution (2 transitions and 4 transversions) in the different crossing over categories (Figure 3.3) using data set 1 only (since this contains the most data). There is no significant heterogeneity in substitution rates across crossing over categories (ANOVA: p = 0.99 for both *D. melanogaster* and *D. simulans*), however, rates of the different types of substitutions do vary significantly. In particular, the G:C → A:T rate is ~2.5-fold higher than any other rate and substitution rates generally are biased towards A:T. If we assume that this pattern is stable in both species through time, we can estimate the expected GC content at equilibrium for INE-1 elements. We did not find any difference in expected GC content at equilibrium among the four recombinational environments in both species (see Supplementary Table 3.1). Combining the data from different recombination environments, the expected GC

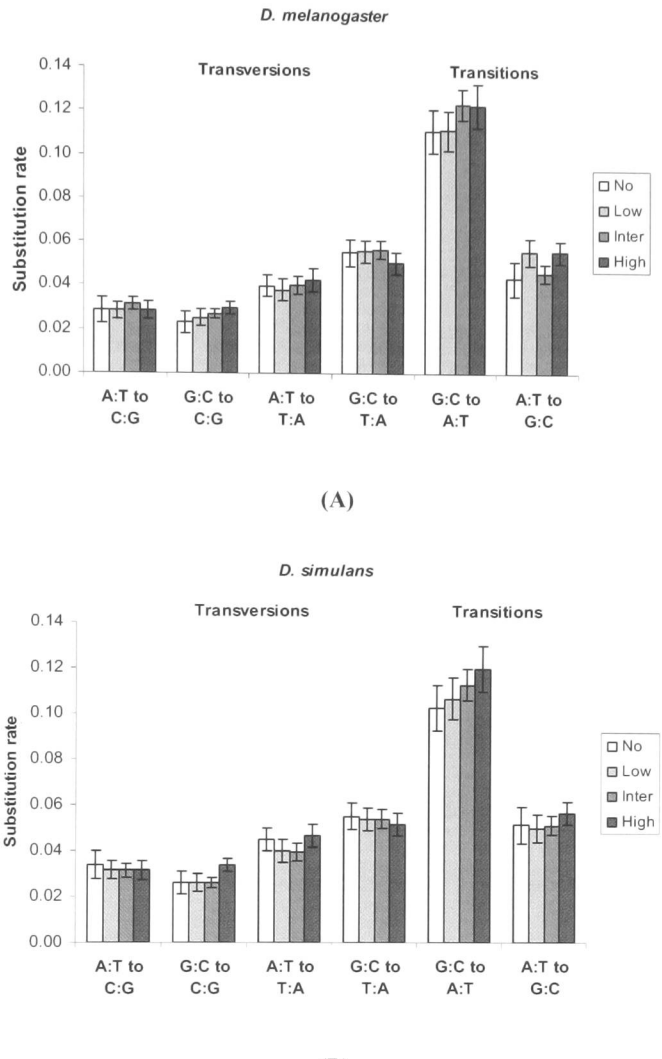


Figure 3.3. - Mean substitution rate (+/– standard error) of six different types of nucleotide substitutions for INE-1 elements between *D. melanogaster* and *D. simulans* from data set 1. Substitution rates were estimated by comparing the ancestral (consensus) sequence to extant INE-1 elements in (A) *D. melanogaster* and in (B) *D. simulans*. The rates are shown for the four different crossing over frequency categories (no, low, intermediate and high crossing over in order).

content at equilibrium is estimated to be 32.3% (95% CI, 29% – 35%) for *D. melanogaster*, and 33.7% (95% CI, 31% – 37%) for *D. simulans*.

Under the neutral model of nucleotide substitutions and a uniform mutation rate, the ratio of transitions (ts) to transversions (tv) is expected to be 1:2 (because there are twice as many possible transversions). Combining elements from the four recombinational categories, we observed 5, 582 transitions and 5, 711 transversions in *D. melanogaster* lineage, and 5, 653 transitions and 5, 812 transversions in *D. simulans* lineage. The ts:tv ratios (1:1.02 in *D. melanogaster*; 1:1.03 in *D. simulans*) are clearly different from 1:2, as expected under a uniform mutation rate (binomial exact test: $p << 0.001$). They are also significantly different from the ts:tv ratio observed for *helena* element in *Drosophila*, 1:1.22 (Petrov and Hartl 1999, binomial exact test: $p < 0.001$). Transitions seem to be more favored in INE-1. This increased transition rate is mainly attributable to G:C → A:T substitutions. However, our estimates of the ts:tv ratio are very close to that estimated for synonymous substitutions between *D. melanogaster* and *D. simulans* in two nuclear genes, *Adhr* and *Adh* (roughly 1:1, Moriyama and Powell 1997). They are also close to the ts:tv ratio (about 1:1) observed in noncoding polymorphisms in *Drosophila* (Moriyama and Powell 1996). This suggests that transition bias may be a general pattern for relatively unconstrained sequences in *Drosophila*.

Heterogeneity of evolution of GC content along INE-1

We have shown that GC content varies dramatically along the INE-1 consensus sequence, and that regions of high ancestral GC content appear to be evolving towards reduced GC content, and vice versa. Thus, it is possible that regions with high ancestral GC content will show different substitution patterns compared to regions with low ancestral GC content. To investigate this, we divided the INE-1 consensus sequence

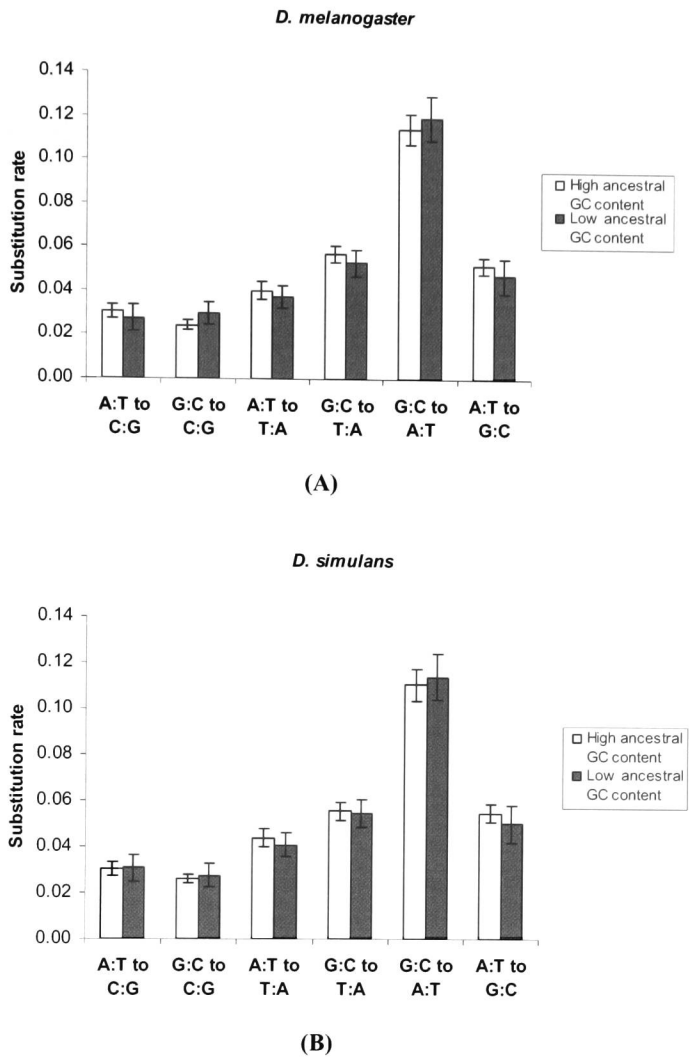


Figure 3.4. - Mean substitution rate (± standard error) of six different types of nucleotide substitutions for INE-1 in (A) *D. melanogaster* and (B) *D. simulans* from data set 1. Here, INE-1 consensus sequence is divided into regions with high ancestral GC content (> 40%) and regions with low ancestral GC content (< 30%). We compare substitutional patterns between the two regions in each species.

(594bp in length) into 12 non-overlapping ~50bp segments and calculated the ancestral GC content within each segment. We arbitrarily split these segments into those with high GC (>40%), moderate GC (>30%, <40%) and low GC (<30%) content (combining

elements from the four recombinational environments). We then examined substitution patterns in the high and low GC content categories in *D. melanogaster* and *D. simulans* for data set 1, using the INE-1 consensus sequence as an outgroup.

This analysis revealed no significant differences between the high and low categories for the six different substitution rates in both species (Figure 3.4). This is what would be expected if the INE-1 sequences were evolving neutrally. However, differences in GC content between these two GC content categories lead to differences in the relative numbers of GC → AT and AT → GC substitutions. For example, in *D. melanogaster*, we observed 1, 981 GC → AT and 1, 176 AT → GC substitutions in the high ancestral GC content category (leading to a reduction in the number of GC bases), whereas in the low GC content category we observed 837 GC → AT and 955 AT → GC substitutions (leading to an increase in the number of GC bases).

3.5. Discussion

We investigated rates and patterns of substitution in three different classes of sites of the *Drosophila* genome, which previous studies have hypothesized to be evolving close to neutrally. We compared evolution of remnants of the most abundant class of TEs in the *D. melanogaster* subgroup (INE-1) with sites within short introns (termed FEI sites) and four-fold degenerate synonymous sites. We found strong evidence for varying rates of substitution between these site classes, between different recombination environments, and between intronic and intergenic sequences. Furthermore, the patterns

are inconsistent between pairs of species that differ in their divergence times. These observations could be attributable to various processes, including time to coalescence in the common ancestor of differences resulting from ancestral polymorphisms, mutation rate variation and selection. These possible explanations in relation to each finding are discussed below.

We found that the mean divergence of four-fold sites is lower than that of FEI sites in all data sets and crossing over categories. This supports previous observations that at least some four-fold degenerate sites are under weak selection for protein translation efficiency (Akashi 1995; McVean and Vieira 2001) and suggests that FEI sites are evolving closer to neutrally (Halligan and Keightley 2006).

We found a positive correlation between rate of crossing over and divergence in FEI/four-fold sites in all three data sets. A similar pattern has previously been observed for the divergence of short introns between *D. melanogaster* and *D. yakuba* (Haddrill et al. 2007). A positive correlation has also been shown in other organisms for four-fold degenerate sites (Lercher and Hurst 2002; Waterston et al. 2002; Hardison et al. 2003; Hellmann et al. 2003). In contrast, Shapiro et al. (2007) found no relationship between recombination rate and synonymous site divergence between *D. melanogaster* and *D. simulans*. The correlation observed for FEI/four-fold sites in our data also contrasts with the results of Begun and Aquadro (1992), who found no correlation between the coefficient of exchange and divergence for 20 genic regions between *D. melanogaster* and *D. simulans*. It also contrasts with results obtained for amino-acid sites in *Drosophila* (Betancourt and Presgraves 2002; Presgraves 2005; Haddrill et al. 2007), in

which divergence was highest in regions lacking recombination. It has been suggested that these latter results are due to variation in the effectiveness of selection, which is expected to be lower in regions with less frequent crossing over due to reduced N_e. This would result in a higher rate of substitution for sites under weak negative selection in these regions. In contrast, the patterns observed for FEI/four-fold sites in our data sets could be explained by weak positive selection on these sequences in regions of no crossing over, such that divergence is inflated above the neutral expectation in regions of high recombination. Alternatively, the patterns observed could be explained by a mutagenic effect of recombination, which has been suggested previously to explain a similar pattern observed for four-fold degenerate sites in mice (Lercher and Hurst 2002; Waterston et al. 2002; Hardison et al. 2003) and in humans (Hellmann et al. 2003). Finally, they may result from differences in the mean coalescence time among recombination rate regions. Total divergence between a pair of species reflects the time since speciation plus the coalescence time of polymorphisms present in the common ancestor. Lack of recombination leads to a lower effective population size because of Hill-Robertson interference and therefore increased rates of genetic drift. Therefore, regions with a low frequency of crossing over would be expected to have the shortest mean coalescence time. All else being equal, this will result in lower apparent divergence for sequences in regions lacking recombination. This interpretation is supported by the fact that the magnitude of the relative differences between recombination rate regions is higher in the *D. simulans vs. D. sechellia* pairwise comparison, for which the ratio of polymorphism in the ancestor to divergence is highest. One should note that the ecology of the three species is very different, thus the

effective population size N_e is likely to be different between them. Indeed, N_e tends to be relatively smaller for *D. sechellia* and larger in *D. simulans*. But we argue that this could only affect the magnitude of the correlation coefficients, but not the general pattern of our results.

We observed markedly different patterns of divergence for INE-1 elements among the three data sets. In regions of high crossing over frequency, INE-1 elements have a lower mean divergence than FEI sites (in all three data sets), and this is significant for data sets 2 ($p = 0.042$) and 3 ($p = 0.01$). This may be a result of weak negative selection acting on some/all INE-1 elements in this category, reducing divergence relative to the neutral expectation. The lack of difference between intermediate, low and no crossing over frequency categories could then be attributed to less effective selection or the lack of any selection on these elements at all. However, the result could also be attributable to a process whereby the coalescence time of polymorphisms present in the common ancestor tends to be shallower for INE-1 elements than for FEI sites. If TE insertions are relatively recent, they may not have been at mutation-drift balance in the common ancestor. This would reduce the expected coalescence time and therefore apparent divergence for these elements. Thus, it is not necessary to invoke selection on INE-1 elements to explain these results. However, we also found a significantly negative correlation between divergence and crossing over frequency in data sets 1 and 2 (not 3). This result is not easily explained by differences in coalescence times, which would predict the opposite relationship, unless INE-1 elements in the high crossing over frequency category tend to be younger. However, INE-1 elements in high crossing over

frequency categories are not more closely related to the consensus sequence (see Figure 3.2), which argues against this interpretation. Therefore, there is some weak support for negative selection on INE-1 elements in highly recombining regions, at least between *D. melanogaster* and *D. simulans*.

We found strong evidence for over-dispersion of substitutions for INE-1 remnants in 9 out of 12 crossing over categories from the three data sets. This could be explained by variation in time to coalescence of ancestral polymorphisms, which is subject to stochastic variance, but is also, at least, partly explained by differences between intergenic and intronic elements. In all three data sets we found that intergenic INE-1 elements evolve faster than intronic ones. This pattern has also recently been observed in rodent TEs (Gaffney and Keightley 2006). Furthermore, this is apparent within crossing over categories, suggesting that it is not an artifact of differences in rates of recombination between intergenic and intronic regions. There are several possible explanations for this result. Firstly, sites in introns may experience a lower mutation rate, *e.g.*, as a result of transcription-coupled repair (TCR), since sites in intronic regions are transcribed, whereas those in intergenic regions are not. This mechanism has been found in bacteria, yeast and mammals (Deaconescu et al. 2006; LePage et al. 2000), but not in *Drosophila*. Secondly, there may be negative selection on transcribed DNA generally. For example, it has been shown that selection in introns of the alcohol dehydrogenase locus (*Adh*) of *Drosophila pseudoobscura* helps maintain secondary structure of pre-mRNA (Kirby et al. 1995). Finally, there may be greater negative

selection acting on intronic INE-1 because, for example, intronic INE-1 elements are more likely to be co-opted for a function than intergenic elements.

Finally, we have investigated the pattern of base composition of INE-1 since insertion by aligning extant remnants with the consensus sequence (after establishing that this is likely to be a reasonable approximation of the true ancestral sequence). We found that substitutions since insertion have tended to be biased towards A and T nucleotides in both *D. melanogaster* and *D. simulans*. Very similar patterns have been observed for those putatively neutrally evolving sites in *Drosophila* (*e.g.*, INE-1 element in Singh et al. (2005a); *helena* element in Petrov and Hartl (1999)). Under the assumption that this is a result of a mutation bias, rather than purifying selection, we estimated the expected equilibrium GC content to be 33.1% (95% CI: 30% – 36%, combining crossing over regions and data from *D. melanogaster* and *D. simulans*). This is consistent with other estimates of the equilibrium GC content for putatively neutrally evolving sites (Petrov and Hartl 1999), and with an estimate of equilibrium GC content for low recombination regions in *Drosophila* (33.0%), based on substitution rates among paralogous copies of INE-1 in *D. melanogaster* (Singh et al. 2005a). However, we did not find any differences in expected GC content at equilibrium among the four recombinational environments, which was shown in a previous study (Singh et al. 2005a). This is possibly caused by a lack of power due to the small sample size of INE-1 elements in regions of no crossing over in our data.

In our data, we have observed differences in interspecies divergence amongst nucleotide site types, amongst regions with different frequencies of crossing over.

Furthermore, in some cases the relative strength and direction of these patterns varies depending on species compared. We also found evidence for over-dispersion of substitutions between INE-1 elements, especially in the close *D. simulans*/*D. sechellia* comparison. However, one should note that applying recombination estimates in *D. melanogaster* to other species may bias our results. We conclude that the majority of the patterns observed can be explained by differences in time to coalescence of polymorphisms in the common ancestor. Furthermore, this process can explain the differences observed between species comparisons, since the magnitude of this effect is expected to be stronger when species are less diverged (*e.g.*, *D. simulans* and *D. sechellia*). Although a mutagenic effect of recombination could produce similar patterns of evolution, here we argue that it is not necessary to invoke this, given that the process described above is likely to be operating. Additionally, a mutagenic effect of recombination could not explain the difference in the magnitude of the observed effect between the two different pairwise comparisons (*D. melanogaster vs. D. simulans* and *D. simulans vs. D. sechellia*). However, we argue that some observations, *i.e.* the faster evolutionary rate for intergenic INE-1 elements than intronic elements, the negative correlation between divergence and frequency of crossing over for INE-1 between *D. melanogaster* and *D. simulans* and faster evolution of FEI sites than for four-fold sites, are difficult to explain by either variation in time to coalescence resulting from polymorphisms in the common ancestor or mutagenic effect of recombination. Instead, we suggest that these observations may result from variation in strength of negative or positive selection among elements.

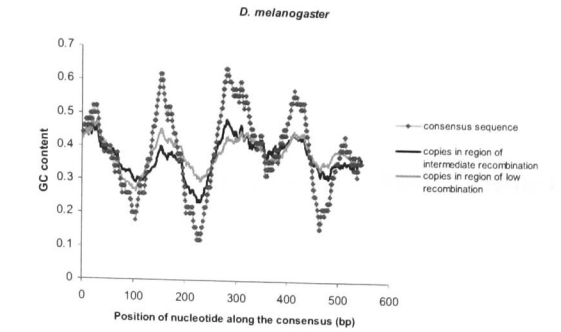
3.6. Acknowledgements

We are grateful to the Genome Sequence Center, WUSTL School of Medicine, the Broad Institute of MIT and Harvard and the Berkeley *Drosophila* Genome Project for providing the genome sequences we analyzed in this study. We also thank Flybase and NCBI for providing genome annotation data. We thank Toby Johnson, Daniel Gaffney and Brian Charlesworth for helpful comments. JW was supported by the Dorothy Hodgkin Postgraduate Studentship Award. Funding for DLH was provided by the Wellcome Trust.

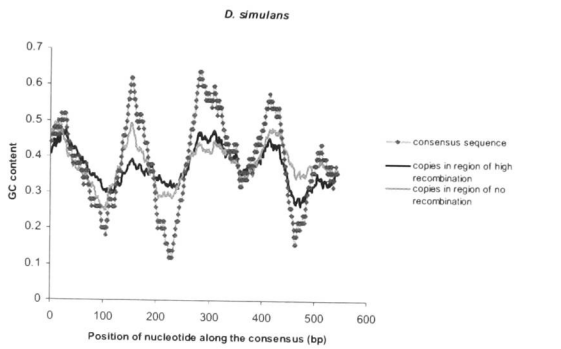
3.7. Supplementary Materials

Supplementary Table 3.1. The expected GC content at equilibrium for INE-1 elements in four recombinational categories from High to No in both species, *D. melanogaster* and *D. simulans*. 95% confidence intervals are shown in parentheses. There are no significant differences in the expected GC content among the four recombinational categories in both species.

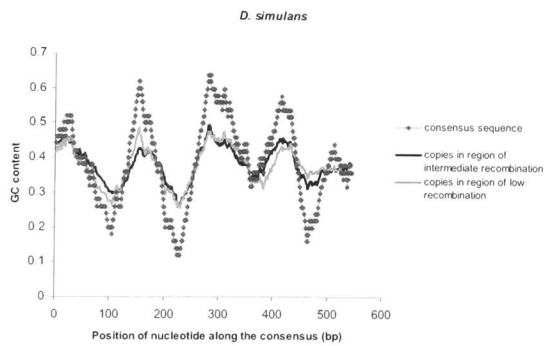
Recombination rate	*D. melanogaster*	*D. simulans*
High	32.8% (29.7% - 35.6%)	33.9% (30.3% - 37.5%)
Intermediate	30.2% (28.1% - 32.4%)	33.2% (30.2% - 36.2%)
Low	33.5% (30.4% - 36.6%)	33.7% (30.5% - 36.0%)
No	30.2% (27.0% - 33.5%)	35.1% (31.4% - 38.5%)



(A)



(B)



(C)

Supplementary Figure 3.1. - Mean GC content of consensus sequence (light gray line with dark dots) and the orthologous INE-1 extant remnants in *Drosophila* for each nucleotide site along the consensus sequence in 50bp sliding windows in data set 1 for *D. melanogaster* and *D. simulans*. Here, we show results for, (A) INE-1 in *D. melanogaster* in regions of intermediate (black lines) and low (gray line) recombination, (B) INE-1 in *D. simulans* in regions of high (black line) and no (gray line) recombination and (C) INE-1 in *D. simulans* in regions of intermediate (black line) and low (gray line) recombination.

Chapter 4.

Transposable Elements Are More Than Just Genomic Parasites: A Case Study in *Drosophila*

Jun Wang*, Peter D. Keightley and Daniel L. Halligan

Istitute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

***Corresponding author**: Jun Wang, e-mail: j.wang-13@sms.ed.ac.uk

JW and DLH initially conceived of and designed the study. JW performed all the data analysis. PDK co-designed and commented on the study. JW wrote the manuscript. PDK and DLH commented on and wrote the manuscript. All authors read and approved the final manuscript.

4.1. Abstract

The evolution of transposable elements (TEs) is one of the central topics of genome evolution. We aim to learn the distribution and rates of evolution of three major TE classes (LTR, LINE retrotransposons and DNA transposons, also known as TIRs) in *Drosophila*. We first investigate the distribution of TEs in the *Drosophila melanogaster* euchromatic genome using a gene-centric approach. Most of our findings are consistent with previous studies. For example, LTR elements appear to outnumber LINE and TIR elements in both intergenic and intronic regions. We also show that TEs constitute ~0.02% of nucleotides of exons (~90% of which are fragments of *roo_I*). We then demonstrate that between *D. melanogaster* and *D. yakuba*, orthologous TIR fragments show a significantly higher mean divergence than orthologous LTRs and LINEs, and of them, *roo_I* fragments appear to be evolving the most slowly, indicating that some LTR sequences/fragments (especially *roo_I*) may have been co-opted by the host, and have become selectively constrained between species. We then investigate TE representation in *Drosophila* promoter regions (1-500bp upstream of start codons), and show that promoter regions have significantly fewer TEs than more distal regions, indicating stronger selection against TE insertions proximal to coding regions. Furthermore, orthologous TEs in promoter regions appear to be evolving more slowly than those in distal regions. This result is mainly due to the contribution of LTRs (mostly *roo_I*) in promoter regions. Our results suggest that TEs may have contributed substantially to gene regulation in *Drosophila*.

4.2. Introduction

Transposable elements (TEs) are one of the major components of the genomes of many organisms. It has long been assumed that TEs are simply genomic parasites with little or no effect on the host genome (Orgel and Crick 1980), and that they can compromise gene function through their deleterious mutations in individuals (Deininger and Batzer 2002). However, it is possible that some TE-derived sequences could become functionally important by donating transcriptional regulatory signals (van de Lagemaat et al. 2003; Thornburg et al. 2006; Fablet et al. 2007) or even by encoding protein sequences (Nekrutenko and Li 2001; Bejerano et al. 2006) that benefit the host. This possibility is supported by a growing body of evidence, especially in species where TEs make up a larger proportion of the genomic content, such as mammals. For example, it has been shown that ~25% of promoter regions in the human genome contain TE-derived sequences (Jordan et al. 2003). Among these TE fragments, some TE classes are more likely than others to carry transcription regulating signals, and have an important potential to contribute to pre-transcriptional gene regulation (Thornburg et al. 2006). Studying the impact of TE insertions on the regulation of gene expression (e.g., as promoter or enhancer sequences for the nearby genes) is therefore a fundamental aspect of understanding how the genome works.

In species where TEs are relatively uncommon, such as *Drosophila*, TE-derived sequences may still play important roles for the host genome evolution (Kazazian 2004). It has been established in the *D. melanogaster* genome that TEs are not randomly distributed, even within regions of high and low TE abundance (Kaminker et al. 2002;

Bergman et al. 2006). The density of TEs increases in the proximal euchromatin, defined as the proximal 2Mb of the assembly of each of the five major chromosome arms (X, 2L, 2R, 3L and 3R) constituting ~10% of the euchromatic sequence (Kaminker et al. 2002). Furthermore, the distribution of three major TE classes (LTR and LINE elements, RNA-mediated; TIR elements, DNA-mediated) tends to differ among chromosomes. LTR elements appear to outnumber the other TE classes on the major chromosome arms, but LINE (also known as non-LTR) and TIR elements have higher numbers relative to LTR elements on chromosome 4 (Kaminker et al. 2002). Although the chromosomal positions of TEs have been clarified since Release 3 sequence of *D. melanogaster*, attention still needs to be paid to how these TE classes are represented in noncoding (intergenic/intronic) and coding regions (in a gene-centric approach), since some TE classes may be overrepresented in some regions due to their different transposition mechanisms or their potential to be co-opted by the host.

The most important prospective of TEs being functional is the ability to potentially contribute their regulatory regions to form new host regulatory sequences. This is supported by evidence from gene promoter regions of mammals (Jordan et al. 2003; Fablet et al. 2007). However, TE's roles in gene regulation may be very lineage-specific (Mariño-Ramírez et al. 2005). First, the proportion of TE-derived sequences in the genome differs dramatically between species, e.g. from more than 50% of the human genome (International Human Genome Sequencing Consortium 2001) to less than 10% of the *Drosophila melanogaster* genome (Quesneville et al. 2005), and up to 90% of the genomes of lilies and wheat (Flavell et al. 1994). Second, the distribution of different

types of TEs in the genome also differs greatly between lineages (Biémont et al. 1997; Kaminker et al. 2002; Thomas et al. 2003; Yang et al. 2006). For example, LINE elements are the most abundant TEs in mammals, accounting for 17% of the human genome (Bannert and Kurth 2004), and LTR elements comprise ~8% of human DNA (International Human Genome Sequencing Consortium 2001). In contrast, LINE and LTR elements account for only 0.9% and 2.7% of the euchromatin of *D. melanogaster*, respectively (Kaminker et al. 2002). Furthermore, the most abundant *Drosophila* TEs are INE-1 (interspersed elements), a family of nonautonomous DNA transposons, comprising more than 1% of the *D. melanogaster* genome. They are believed to result from an ancient transposition event (~5-10 MYA) in the *D. melanogaster* species complex (Kapitonov and Jurka 2003; Singh et al. 2005; Bergman et al. 2006; Slawson et al. 2006). INE-1 elements are absent in the human genome. However, human genomes contain many copies of other short interspersed elements (SINEs), such as the *Alu* elements (Deininger et al. 2003), which *Drosophila* genomes lack. When these findings are considered with respect to the influence of TEs on gene regulation, patterns of TEs serving as regulatory sequences in human (Jordan et al. 2003; Thornburg et al. 2006) may deviate from those observed in *Drosophila*.

Here, we first investigate the distribution and representation of major TE classes in intergenic, intronic and exonic regions of the *D. melanogaster* euchromatic genome. We also compare the rate of evolution of fragments among TE classes, since some TE fragments may be more likely to be involved in host gene regulation and have become selectively constrained between species. This is done by calculating mean divergence of

orthologous TE elements between *D. melanogaster* and *D. yakuba*. We chose this species pair (rather than closer species, *D. melanogaster* and *D. simulans*) because they should show relatively larger differences in mean divergence between TE classes (if they exist). Meanwhile, we also consider the effect of crossing over rates, since they have an impact on the efficacy of natural selection (Hill and Robertson 1966; Kliman and Hey 1993; Hey and Kliman 2002; Haddrill et al. 2007) and impact on the coalescence time in the common ancestor (Hudson 1991, p. 1-44; Takahata et al. 1995; Yang 1997; Wang et al. 2007). Secondly, we investigate TE insertions in promoter regions of *Drosophila*, and compare their distribution and divergence to distal promoter regions immediately adjacent to the 5' end of promoter regions, since their genetic background should be very similar. We test whether promoter regions have significantly fewer TE insertions than the control regions. We also investigate TE insertions in 3' flanking regions of genes, since insertions may be also selectively constrained here. Finally, we test whether TE-derived sequences in promoter regions/proximal 3' regions evolve more slowly that those in more distal regions based on the comparative analysis of orthologous TE fragments between *D. melanogaster* and *D. yakuba*. This rests on the assumption that TEs in promoter regions are more likely to be functional, and will thus evolve more slowly than non-functional TEs.

4.3. **Materials and Methods**

Definition of some genomic regions for promoter regions investigation

For each gene, we defined three genomic regions: (1) Proximal 5' region (promoter region), which ranges from position 500bp to 1bp upstream from the transcription start site of coding sequences (CDS). (2) Distal 5' region, which ranges from position 1,000bp to 501bp upstream from the transcription start site of CDS. (3) Proximal 3' region, which ranges from position 1bp to 500bp downstream from the transcription end site of CDS. Each region is 500bp in length. The Proximal 5' region contains most (almost all) important promoter sequences, such as core promoters (~35bp upstream from the transcription start site) and proximal promoters (~250bp upstream), whereas the distal 5' region may contain some distal promoters, although they are unlikely to be as important as core and proximal promoters (see review by Smale and Kadonaga 2003). The proximal 3' region also contains some elements that function as binding sites for proteins or miRNAs (e.g., enhancers).

Compilation of sequence data

We used the Ensembl GenBank database file for the genome sequence data of *D. melanogaster* (http://www.ensembl.org/Drosophila_melanogaster/index.html). The euchromatic sequence is based on BDGP assembly release 4, whereas the annotations of the euchromatic regions displayed in Ensembl are based on data imported from FlyBase (release 4.3, dated 2006/01/30). This retrieved a total of 14,387 euchromatic gene annotations. Any poorly annotated genes and RNA genes were excluded (this was achieved by examining the FlyBase synopsis report for each gene, and excluding genes that were based on BLASTX data or gene prediction data only). We also excluded ~1000 genes that are nested within the other genes. This left us a total of 12,474

euchromatic nonnested genes. We then extracted intergenic sequences (i.e., from the translation end site of the previous gene to the translation start site of the present gene) and coding sequences (CDS) from these genes. For genes that are alternatively spliced, only the form with the longest CDS was used. For each intergenic sequence and coding sequence, we identified TE fragments residing within them using RepeatMasker (Http://www.repeatmasker.org). The following parameters were used for this search: "cross_match" as the search engine; "nolow" to not mask low complexity DNA or simple repeats; "norna" to not mask small RNA (pseudo) genes; "no_is" to skip bacterial insertion element check. In addition to the parameters selected from the program, our analysis identified a TE insertion as a sequence of at least 80bp in length that also possessed at least 75% identity to the canonical (consensus) sequence in the RepeatMasker library database. These stringent parameters were set to avoid spurious results. Similar strategies have been used to identify TEs in the *Bos taurus* genome (Almeida et al. 2007).

To further test the reliability of RepeatMasker (i.e., it does not just pick up TE matches by random in the genome), we also ran RepeatMasker on *C. elegans* and honey bee genomes using the *D. melanogaster* TEs as query sequences. The same parameters were used in RepeatMasker, and the same criteria for TE identification discussed above were employed. TEs identified in this way only comprised ~0.02% of the total genomic content in both *C. elegans* and honey bee. These figures were much lower than the percentage of TEs in the *D. melanogaster* genome, ~5.5%. Furthermore, most of the TEs identified in *C. elegans* and honey bee were from very different TE families. For

example, most of the TE fragments identified in *C. elegans* are DNA/HAT and DNA/Tc1 elements. TEs that are abundant in *D. melanogaster* are very rare or even absent in both *C. elegans* and honey bee. We thus believed that RepeatMasker is a reliable tool for identifying TEs given TE consensus sequences (especially after using the criteria for TE identification, 80bp in length and 75% in sequence identity).

We categorized TE fragments identified in coding sequences into those in intronic regions and those in exonic regions. For any TEs that overlap an exon and an intron, we split them into fragments in exons and fragments in introns. We calculated the proportion of TE-derived sequences in each intergenic, intronic and exonic sequence. We then calculated the mean proportion of TE-derived sequences for all intergenic, intronic and exonic regions, respectively. 95% confidence intervals for the mean proportion were obtained by bootstrapping by each sequence for the three genomic categories.

We also compared TE fragments in promoter regions with those in distal promoter regions, and also with those in proximal 3' regions. Because we tried to avoid the case that the proximal 3' region of the previous gene overlaps with the distal 5' region of the present gene, we only used genes whose flanking sequences are longer than 1,500bp in length in both 5' and 3' regions. We then extracted 1,500 nucleotides upstream/downstream from the transcription start/end site of CDS of available genes (6,763 nonnested euchromatic gene annotations in our final dataset for the investigation of TEs in promoter regions). The 5' and 3' flanking sequences of length 1,500bp were then scanned for occurrence of TEs using RepeatMasker. Note that TE fragments in

those regions are a small subset of TE fragments in the whole intergenic regions. After TEs had been masked, each 1,500bp 5' flanking sequence was split into a 500bp promoter region and a 500bp distal 5' region as discussed above. Proximal 3' regions were extracted from 6,763 3' flanking sequences of length 1,500bp. Within these three regions, we counted the number of nucleotides that are derived from TEs and the number of these regions that contain at least one TE-derived sequence or fragment. Here we excluded microsatellites and TEs of unknown family.

Analysis of gene overrepresentation

For those 12,474 genes in euchromatin, it is possible that genes associated with some particular functional or biochemical process (i.e., Gene Ontology (GO) terms information), or clustered in a region of the genome are more likely to have TE-derived sequences in their coding sequences, or even exonic sequences. We carried out the overrepresentation test using GeneMerge (Castillo-Davis and Hartl 2003), which is a program that returns functional and categorical genomic data for a given set of genes and provides statistical rank scores for over-representation of particular functions or categories in the dataset.

Alignment of orthologous TEs and noncoding sequences between D. melanogaster and D. yakuba

For the whole-genome alignment between *D. melanogaster* and *D. yakuba*, we used that generated by Assembly/Alignment/Annotation of 12 *Drosophila* Genomes Project (AAA 12 genomes alignments, *Drosophila* 12 Genomes Consortium 2007), in which

multiple whole-genome alignments of *Drosophila* species were generated by Mercator (an orthology mapping program, http://www.biostat.wisc.edu/~cdewey/mercator) and MAVID (a multiple alignment program, Bray and Pachter 2004). The whole-genome alignments of D. *melanogaster* (BDGP assembly release 4) and *D. yakuba* (Washington University Release DroYak2.1 reconciled with Arachne/Celera assemblies) were constructed with 112 Mercator assembled orthologous contigs with the average length of 1,063,528bp for sequences from *D. melanogaster*. The coordinates for each contig were specified with positions in the *D. melanogaster* genome sequence.

Since we knew the positions of all TE sequences/fragments in the *D. melanogaster* genome, we can identify the orthologous TEs in *D. yakuba* by searching for the corresponding sequences to *D. melanogaster* TE sequences according to the whole-genome orthology alignments. We tried to eliminate the possible noise in our alignments introduced by recent transpositional events of LTR in *D. melanogaster* (Bergman and Bensasson 2007) and by those of INE-1 in *D. yakuba* (Yang et al. 2006), by excluding any TE alignment whose 100bp flanking regions contain TE fragments that are from the same family, since these TEs may have been newly inserted within or next to older remnants (from the same family) and would compromise alignments for real orthologous TEs. We also checked the orthology of TEs by examining alignments of flanking sequences with 50bp in length at both 5' and 3' ends of TEs manually. We randomly selected 50 genes and found that the alignment quality was reasonably good for flanking sequences (i.e., mean divergence for flanking sequences of these 50 genes was ~0.25, and none of divergences of these flanking sequences exceeded 0.50). We then extracted

the pairwise alignments of orthologous TEs from the whole-genome alignments of *D. melanogaster* and *D. yakuba*. These alignments were then refined using MCALIGN2, a pairwise alignment method with an explicit model for indel evolution (Wang et al. 2006). We excluded all alignments with fewer than 80 valid bases (i.e., A, G, C or T) in any species. We also masked sections within TE alignments in which there were short lengths of bases surrounded by long gaps, and/or removed alignments whose divergence was >0.50 (approximately twice the divergence between *D. melanogaster* and *D. yakuba*). In the end, we extracted ~2,500 orthologous TE elements/fragments in intergenic regions, ~500 in intronic regions and 42 in exonic regions. We show the procedure of identifying TEs and extracting orthologous elements/fragments in Supplementary Figure 4.1. We also show some cases of orthologous TE alignments between *D. melanogaster* and *D. yakuba* in Supplementary Figure 4.2.

We also extracted whole orthologous noncoding (intergenic/intronic) sequence alignments between *D. melanogaster* and *D. yakuba* from the AAA 12 genomes alignments. Any noncoding sequence in *D. melanogaster* whose orthologous sequence did not exist in *D. yakuba* was deleted. This left us 9,207 and 35,782 orthologous intergenic and intronic sequences in *D. melanogaster* and *D. yakuba*, respectively. All divergence estimates were corrected for multiple hits (Kimura 1980). Mean divergence was calculated on a per site basis, and 95% confidence intervals were calculated by bootstrapping by TE element or noncoding sequence.

Using the same strategies, we also extracted 130, 187 and 157 orthologous TE fragments in promoter regions, distal 5' regions and proximal 3' regions from pairwise

orthology alignments between *D. melanogaster* and *D. yakuba*, respectively. These TE alignments cover 13,306 nucleotides in promoter regions of *D. melanogaster* (~34% of TE sequences identified in these regions). For distal 5' regions and proximal 3' regions, 25,963 and 14,108 nucleotides in *D. melanogaster* are covered, making up 31% and 30% of the total TE- derived nucleotides in each region, respectively. Thus, more than 65% of TE-derived sequences from these three regions in *D. melanogaster* do not have any ortholog in *D. yakuba*, or they (orthologs) tend to be too diverged to be recovered. Most of these orthologous TEs are LTR and INE-1 fragments, especially in promoter and proximal 3' regions. Within LTR orthologous fragments, *roo_I* fragments are the most abundant (~70%).

4.4. Results and Discussion

Representation of TEs in intergenic and intronic regions of euchromatin

It has been shown that the centromere-proximal regions of each major chromosome arm have a much higher density of TEs (~4.7 times on average) than elsewhere, and major TE classes tend to have different distributions between chromosomes (Kaminker et al. 2002). Here, we aimed to investigate the distribution of TE-derived sequences in euchromatin in a gene-centric approach, by comparing distributions of TE-derived sequences amongst intergenic, intronic and exonic regions across the genome.

Consistent with the previous study (Kaminker et al. 2002), we found that the number of LTR elements/fragments is much higher than that of LINE and TIR elements/fragments

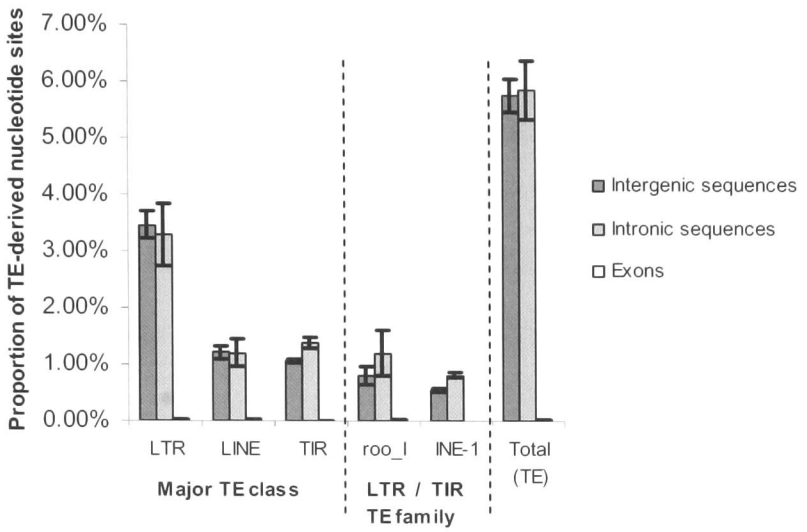


Figure 4.1. - Proportion of LTR, LINE and TIR derived nucleotide sites in intergenic, intronic and exonic regions in *D. melanogaster*. Here, we also show proportions of INE-1 (a type of TIR elements) and *roo_I* (a type of LTR elements) separately due to their huge abundance. 95% CIs are shown by bars.

in both intergenic and intronic regions (i.e., >2 times the number of LINE and TIR fragments, Figure 4.1). Major TE classes are very similarly distributed between intergenic and intronic regions of euchromatin (Wilcoxon test, $p = 0.82$), except that the proportion of TIR-derived sequences in intronic regions is significantly higher than that in intergenic regions ($p < 0.0001$). This result is mainly explained by the higher proportion of INE-1 elements/fragments in intronic sequences (Figure 4.1). The possible explanation for this could be that intergenic sequences are, on average, much longer than introns, and it has been demonstrated that long noncoding sequences are selectively more constrained than short ones (Haddrill et al. 2005; Halligan and Keightley 2006). Thus, many INE-1 insertions in long noncoding sequences may have been removed by natural selection and did not become fixed in the population. This would result in

relatively more fixed INE-1 elements in short noncoding sequences, giving rise to the bias towards introns. More INE-1 insertions in introns than in intergenic sequences have also been observed in a recent study of evolutionary dynamics of INE-1 in 12 *Drosophila* species (Yang and Barbash 2008). In total, TE-derived sequnces account for ~6% of the nuclear content in noncoding (intergenic/intronic) regions of euchromatin in our data, consistent with the previous estimate (Quesneville et al. 2005).

About 40% of LTR elements/fragments in intergenic and intronic regions are *roo_I* fragments, a *Pao* family retrotransposon. Within all 686 and 154 *roo_I* TE sequences/fragments in intergenic and intronic DNA sequences, we recovered 50 (~7.3%) and 19 (~12.3%) elements/fragments whose length is more than 90% of the canonical length (>8000bp in length), respectively. Most of the *roo_I* fragments in noncoding regions (~80%) are therefore short ancient remnants (with length of less than 10% of the canonical length), assuming long fragments are from more recent transpositions. Some of these short fragments may result from element nesting (i.e., TEs have inserted within another element), and LTR elments are believed to be nested more often than either LINE or TIR elements (Kaminker et al. 2002; Bergman et al. 2006). Furthermore, although it has been demonstrated that the majority of LTR families show a pseudogene-like mode of evolution, this has not been established as a general pattern for all LTR and LINE families (Bergman and Bensasson 2007). In fact, Bergman and Bensasson (2007) have found evidence that purifying selection has operated on the terminal branch substitutions inferred to occur since retrotransposon insertion for several

LTR families. Thus, it is possible that some *roo_I* elements/fragments may have been constrained and co-opted for a function by the host.

Gypsy family retrotransposons (e.g. *gypsy*, *gypsy2*, *gypsy8*, *gypsy12*, *stalker*, *stalker2*, *burdock*, *invader2*) are also quite common in intergenic and intronic regions, but most of them comprise only ~1% of the total LTRs each. More than 50% of TIR elements are INE-1 elements/fragments. The distribution of length of these elements/fragments is, however, not as skewed as that of *roo_I* elements/fragments. It has been shown that INE-1 elements/fragments are among the fastest evolving sites in the *Drosophila* genome (Singh et al. 2005; Wang et al. 2007).

Representation of TEs in exonic regions of euchromatin

It has been suggested that TEs may also contribute their coding potential to the host gene (Nekrutenko and Li 2001). We identified TE-derived sequences in exonic regions by first searching for overlapping sections between positions of identified TEs and positions of exons for all 12,474 euchromatic genes in our dataset. To make sure that these identified TEs in exons were real protein-coding sequences (due to possible incomplete/overlapping ORFs), we revisited their properties using the UCSC *D. melanogaster* Genome Browser Gateway (http://genome.ucsc.edu/cgi-bin/hgGateway). We only included those elements that (1) also appeared in exons or crossed intron/exon boundaries according to the latest version of FlyBase annotations, and (2) were also supported by *D. melanogaster* ESTs that have been spliced.

These stringent criteria resulted in just over 30 genes (~0.26%) that contain TE-derived sequences in their exons. TEs only accounted for ~0.021% of the nuclear content in exonic regions. The majority of these TE-derived sequences (~90%) are *roo_I* fragments, whereas the others are just simple repeats and some TART fragments (located in telomeres). The *roo_I* fragments in exonic regions are mostly short fragments, spanning from position ~500bp to ~1000bp within the canonical sequence (>8000bp in length). LTR retrotransposons usually have long terminal repeats and slightly overlapping ORFs for *gag*, *prt*, *pol* and *env* genes that encode a viral particle coat (GAG) and a reverse transcriptase (RT), ribonuclease H (RH), and integrase (IN) to provide enzymatic activities for making cDNA from RNA and inserting it back into the genome (Kazazian 2004). The *roo_I* fragments in exonic regions (length ranging from 80bp to >400bp) appear to be close to the transcription start sites of the canonical sequence, and sites in this region may have more ability for gene regulation and/or protein-coding potential for the host genes than those elsewhere.

If these TE-derived sequences are not real protein-coding sequences subject to selective constraint, the ratio of point substitution across codon positions of the host exons in these TE fragments is expected to be $\approx$ 1:1:1 for first:second:third codon positions. We carried out this test by extracting alignments of exonic TEs between *D. melanogaster* and *D. yakuba* from AAA 12 *Drosophila* genomes alignments available in the UCSC Genome Browser. Annotation for exons in *D. melanogaster* was obtained from FlyBase (release 4.3). We found that the mean divergence of the third codon position was significantly higher than that of the first and second codon positions for those TE-

Table 4.1. Results of gene Ontology (GO) terms overrepresentation test for those 32 genes containing TE-derived sequences within their exons in terms of molecular function, biological process, cellular component and also chromosome position using GeneMerge.

GMRG_Term	Pop_frac	Study_frac	Raw_es	e-score	Description
Molecular function					
GO:0003700	0.028085	8/32	2.09E-06	7.30E-05	transcription factor activity
GO:0043565	0.0142414	4/32	0.001047	0.036658	sequence-specific DNA binding
GO:0003702	0.0081152	3/32	0.002169	0.075926	RNA polymerase II transcription factor activity
Biological process					
GO:0007417	0.0046941	3/32	0.000442	0.047774	central nervous system development
GO:0030154	0.0007956	2/32	0.000279	0.030134	cell differentiation
GO:0045449	0.0121728	4/32	0.000582	0.062875	regulation of transcription
GO:0006355	0.0167078	5/32	0.000173	0.018694	regulation of transcription, DNA-dependent
GO:0006357	0.0355637	7/32	0.000107	0.011527	regulation of transcription from RNA polymerase II promoter
Cellular component					
GO:0005634	0.0755828	10/32	8.05E-05	0.000805	Nucleus

GMRG_Term: the Gene Ontology (GO) terms, which can be found in FlyBase.
Pop_frac: fraction of contributing genes for such GO term in the population genes (12, 474 genes in total).
Study_frac: fraction of contributing genes for this GO term in the study genes (214 genes).
Raw_es: raw e-score for the significance test.
e-score: adjusted e-score for the significance test by GeneMerge.
Description: description of the corresponding GO terms

derived sequences in exons ($p < 0.0001$), with the mean divergence as, codon position 1: 0.026 [95%CI: 0.015 - 0.036], codon position 2: 0.032 [95%CI: 0.016 – 0.048] and codon position 3: 0.126 [95%CI: 0.096 – 0.156]. We also calculated the mean divergence of the three codon positions for randomly selected non-TE derived sites of exons where TEs were inserted as codon position 1: 0.027 [95%CI: 0.010-0.043], codon position 2: 0.042 [95%CI: 0.025 – 0.060], and codon position 3: 0.116 [95%CI: 0.092 – 0.140]. Patterns of codon position substitutions for exonic TEs are then very similar to those of non-TE derived sites from the same exons (Wilcoxon test, $p = 0.75$). This result

suggests that TE-derived sequences in exons appear to be real protein-coding sequences for the host gene, rather than false positives.

To further test the role played by these TE fragments in exons, we carried out the Gene Ontology (GO) terms overrepresentation test for those 32 genes containing TE-derived sequences within their exons in terms of molecular function, biological process and cellular component using GeneMerge (Castillo-Davis and Hartl 2003). We list significant results in Table 4.1. It is clearly shown that genes associated with transcription factor activity and sequence-specific DNA binding are more likely to recruit TE fragments, mostly *roo_I*, as parts of their exons. It is believed that LTR elements carry more transcription regulating signals than LINE/TIR elements (van der lagemaat et al. 2003; Thornburg et al. 2006); they are therefore rarer in gene regulatory regions and protein-coding regions, probably because a high number of regulatory signals are more likely to alter gene expression to a greater extent and have deleterious effects (Thornburg et al. 2006; Fablet et al. 2007). However, LTR fragments (mainly *roo_I*) strongly dominate TEs in exonic regions in our data. This may appear to be an indication of positive selection for some *roo_I* fragments being recruited for a function when urgent recruitment is needed to cope with the changes of genetic environment (Ludwig et al. 2000; Fablet et al. 2007).

It is also noteworthy that we found genes associated with central nervous system development and cell differentiation are more likely to have TE-derived sequences in their exonic sequences (Table 4.1). Recruitment of TE-derived sequences in coding regions may be important for the neural system development. This is supported by

previous observations that most of the genes exhibiting TE regulatory regions are involved in functions such as the stress response and immunity in *Drosophila*, and response to external stimuli in *Escherichia coli* (Arnault and Dufournel 1994; Rocha et al. 2002; Fablet et al. 2006). Creation of new genetic variability from the increase in TE mobility thus can be useful in the face of stressful conditions (Capy et al. 2000), e.g., for central nervous system development. We also found that genes associated with regulation of transcription (e.g., DNA-dependent or from RNA polymerase II promoter) are more likely to recruit TE sequences/fragments. TE-derived sequences may serve as a direct source of transcription regulating signals for the host (Thornburg et al. 2006), or have impact on the regulation of transcription through their "epigenetic" effect (Biémont and Vieira 2006; Slotkin and Martienssen 2007). We also found that retrotransposons (mainly *roo_I*) are more likely to localize in the nucleus. However, this localization may appear to be element-specific. This is suggested by the observation that element-specific localization of *Drosophila* retrotransposon *gag* proteins occurs in both nucleus and cytoplasm (Rashkova et al. 2002).

Density of TEs within intergenic and intronic sequences of euchromatin

In this section, we investigated how TE fragments reside within intergenic and intronic regions. We knew the start and end coordinates for each intergenic and intronic sequence within the *D. melanogaster* genome. We divided each intergenic and intronic sequence into ten non-overlapping length percentage portion sections (e.g. 0%-10%, 10%-20% of the whole length) from the start to the end. Note that this percentage section order (i.e., 0%-10%, 10%-20% to 90%-100%) is not according to any protein

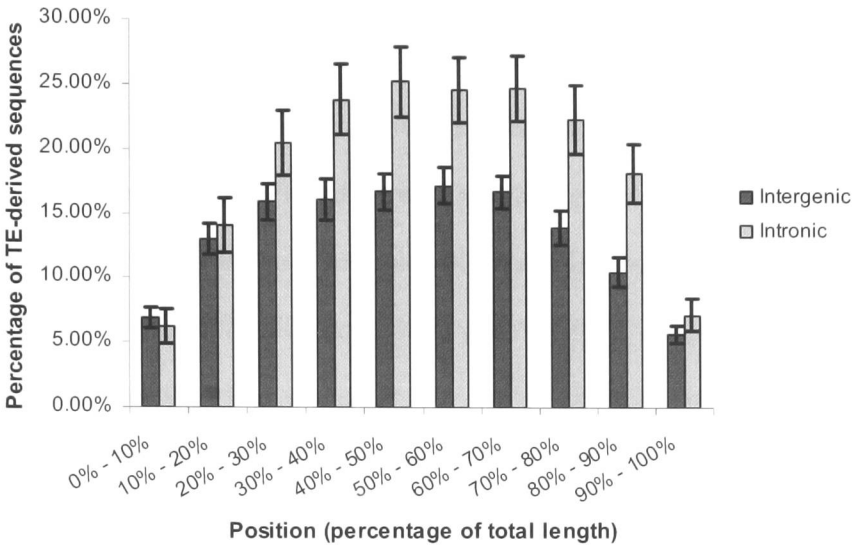


Figure 4.2. - TE density within each percentage portion category. The whole intergenic/intronic sequences were divided into 10 nonoverlapping percentage sections. Within each section, the proportion of TE-derived nucleotide sites was investigated. 95% CIs are shown by bars.

translation order, but only the coordinate order within the genome. Within each percentage section we investigated the proportion of TE-derived nucleotide sites. We carried out this test on 1,879 intergenic and 754 intronic sequences where TE fragments were located. Other intergenic and intronic sequences that do not contain TEs were excluded from this analysis. As expected, in both intergenic and intronic regions, there are significantly more TE-derived sequences in the middle sections (e.g. 40%-50%, 50%-60%) than in sections that are closer to the edges (Figure 4.2). This is probably because the edges of intergenic and intronic sequences are more likely to be involved in gene regulation of their adjacent genes, and TE insertions within such regions may cause disruption for such functions (e.g. transcription initiation or termination), and will

eventually be deleted by natural selection (Biémont et al. 1997; Charlesworth et al. 1997; Duret et al. 2000; Jordan et al. 2003; Thornburg et al. 2006). There is also a clear difference in the percentage of TE-derived sequences between section 10%-20% and 80%-90%, but the direction of this difference is opposite between intergenic and intronic sequences (Figure 4.2). Since our definition of percentage sections is merely coordinate-based, the reason behind this difference (e.g., from the point view of gene regulation for the host gene) is unclear. This difference may just be the artifact of our method dividing the sequences. We also found that the density of TE-derived sequences is significantly higher in intronic regions than in intergenic regions for percentage sections from 30%-40% to 80%-90% ($p < 0.001$, Figure 4.2). This again may result from the fact that intronic sequences are, on average, much shorter than intergenic sequences, and selective constraints tend to be lower in short noncoding sequences (Haddrill et al. 2005; Halligan and Keightley 2006).

Divergence of orthologous TEs in intergenic, intronic and exonic regions of euchromatin

We also wanted to learn whether there are any differences in evolutionary rates between major TE classes, and also, we wanted to test whether TE fragments from different genomic regions (i.e., intergenic, intronic regions and exons) have different divergences. We carried out these tests by calculating interspecies divergence of orthologous TE fragments between *D. melanogaster* and *D. yakuba* across the euchromatic genome.

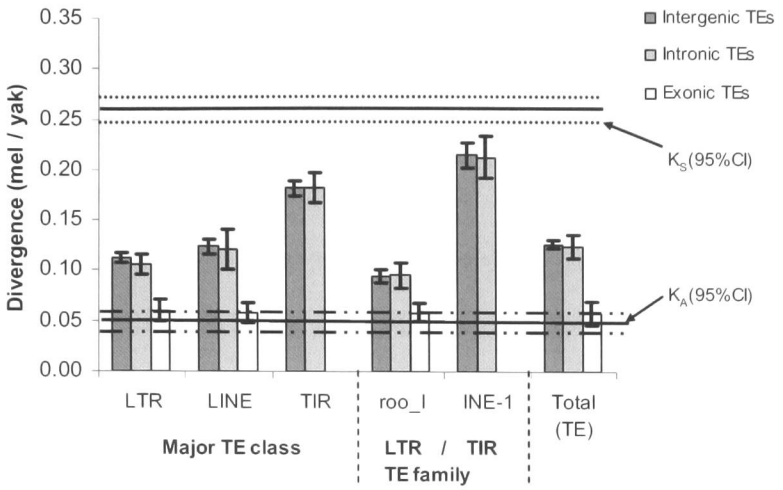


Figure 4.3. - Mean divergence of orthologous TEs of *D. melanogaster* and *D. yakuba* in intergenic, intronic and exonic regions. We show results for three major TE classes (LTR, LINE and TIR). We also show results for two abundant TEs, INE-1 and *roo_I* elements separately. We then show the mean divergence for all TEs in the three regions. 95% CIs are shown by bars. We also show the mean divergence for synonymous (K_S) and nonsynonymous (K_A) sites (Haddrill et al. 2007) by solid lines as comparisons. 95% CI are shown by dashed lines.

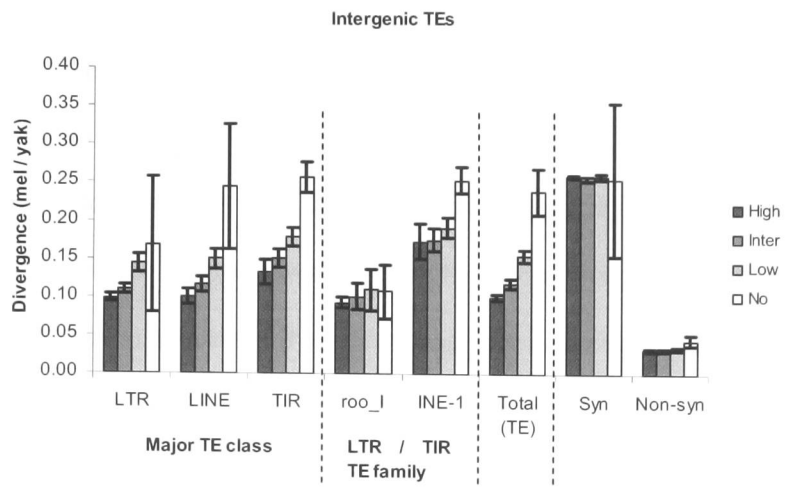
As shown in Figure 4.3, we found that orthologous LTR and LINE fragments have very similar divergences in both noncoding and coding regions. However, their divergences are significantly lower than that of TIR fragments ($p < 0.001$) in both intergenic and intronic regions. We failed to recover any orthologous TIR fragments in exonic regions. Note that we also did not find any orthologous LTR fragments on chromosome 4. Within LTR fragments, *roo_I* fragments tend to be evolving even more slowly than other LTR fragments. This is significant for intergenic regions ($p < 0.05$), but not for intronic regions ($p = 0.46$). We also show the mean divergence for synonymous (K_S) and nonsynonymous (K_A) sites between *D. melanogaster* and *D. yakuba*, calculated by Haddrill et al. (2007). We found that all orthologous LTR, LINE

and TIR fragments had significantly lower mean divergence than synonymous sites (Figure 4.3). This result suggests that TE fragments may be under substantial constraints, most of which affect LTR/LINE fragments. Within TIR elements/fragments, INE-1 fragments appear to be evolving faster than the average of the other element families among TIRs (significantly in intergenic regions, $p < 0.05$). However, they are still evolving significantly more slowly than synonymous sites (Figure 4.3), suggesting that some INE-1 fragments may still be selectively constrained. This result contradicts the finding that INE-1 elements are the fastest evolving sites between *D. melanogaster* and *D. simulans*, even evolving marginally faster than synonymous sites (Wang et al. 2007).

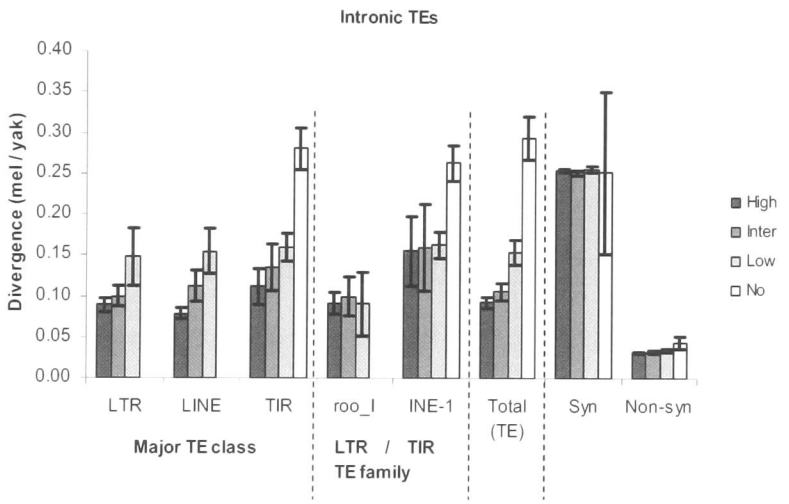
As expected, the total mean divergence for all TE fragments in noncoding (intergenic/intronic) regions is significantly higher than that for all TEs in exonic regions ($p < 0.0001$, Figure 4.3). This is because sites within exonic regions are, on average, more selectively constrained than those in noncoding regions, although there has been mounting evidence that some sites in noncoding regions may be as constrained as sites in coding regions (selection coefficients may be very different, Halligan et al. 2004; Siepel et al. 2005; Halligan and Keigthley 2006).

Divergence of ortholgous TEs from different crossing over environments

It has been shown that the density of TIR elements is significantly negatively correlated with recombination rate, but this tendency is not clear for LTR and LINE elements (Rizzon et al. 2002). It is therefore possible that our result of TIR elements/fragments



(A)



(B)

Figure 4.4. - Mean divergence of orthologous TEs of *D. melanogaster* and *D. yakuba* from different crossing over regions (high, intermediate, low frequency of crossing over and no crossing over) for intergenic (A) and intronic (B) regions. The mean divergence was shown for LTR, LINE and DNA elements, in which *roo_I* and INE-1 were also shown separately. We also show the mean divergence for synonymous (Syn) and nonsynonymous (Non-sys) sites in each crossing over region (Haddrill et al. 2007). 95% CIs are shown by bars.

evolving faster than LTR and LINE elements/fragments is strongly influenced by the distribution of TEs in different crossing over regions for those TE classes. There are many more orthologous TIR elements/fragments in regions with low rate of crossing over than orthologous LTR and LINE elements/fragments. The difference in the coalescence time in the common ancestor and/or in the efficacy of selection on some TE fragments (if it exists) between different crossing over regions will affect our divergence test (Haddrill et al. 2007; Wang et al. 2007). We therefore divided orthologous TE fragments from intergenic and intronic regions of *D. melanogaster* and *D. yakuba* into four crossing over frequency categories (high, intermediate, low frequencies of crossing over and no crossing over) based on the cytologic map location described in Charlesworth (1996), and tested the difference in mean divergence within and among categories.

We found that orthologous TIR fragments evolve relatively faster than LTR and LINE fragments, and this seems to be a general tendency for almost all categories of crossing over in both intergenic and intronic regions (Figure 4.4). The difference in mean divergence between TIR fragments and LTR, LINE fragments appears to be significant for intergenic TEs in regions with high, intermediate and low rate of crossing over ($p < 0.01$), but not in regions of no crossing over ($p = 0.48$ between intergenic TIR and LTR fragments, $p = 0.76$ between intergenic TIR and LINE fragments, Figure 4.4.A). For intronic TEs this difference appears to be significant only in regions with high rate of crossing over ($p < 0.05$, Figure 4.4.B). If we assume that all three types of TE fragments have similar ancestral population sizes (i.e., the coalescence time in the common

ancestor contributes on the total divergence time similarly for all three major TE classes), our results suggest that orthologous LTR and LINE fragments may have been more selectively constrained than orthologous TIR fragments throughout the time since the split of *D. melanogaster* and *D. yakuba*, especially in regions with high rate of crossing over, because the selective efficacy tends to be strongest here. This difference becomes less clear in regions that lack crossing over, possibly due to a lack of data (Figure 4.4).

It is noteworthy that within each TE class, we found that the mean divergence appeared to be negatively correlated with rate of crossing over. This pattern became clearer after we grouped different TE fragments from the same crossing over environment together (total TEs, Figure 4.4). This is consistent with a previous study of orthologous INE-1 elements between *D. melanogaster* and *D. simulans* (Wang et al. 2007). We argued this correlation may result from the some form of selection operating on TE elements/fragments, which has higher efficacy in regions with high frequency of crossing over. Our results appear to support this argument, and indicate further that selection on TEs is possibly prevalent for all major TE classes, and the strength of the selection, however, may be stronger for LTR elements/fragments, possibly due to more transcriptional regulatory signals in them.

We also compared the mean divergence of all TEs with that of synonymous and nonsynonymous sites in each recombinational environment shown by Haddrill et al. (2007). We found that the mean divergence of all TEs was significantly lower than that of synonymous sites (in high, intermediate and low recombinational regions), but

significantly higher than that of nonsynonymous sites (in all recombinational regions) (Figure 4.4). The difference in mean divergence between TEs and synonymous sites increases with the rate of crossing over. This, again, indicates that there may be some amount of selection operating on the evolution of TEs (LTR especially) since the divergence between *D. melanogaster* and *D. yakuba*, at least for those in regions of high/intermediate frequencies of crossing over.

Relationship between TE insertions and their host noncoding sequences

It is generally agreed that euchromatic TE insertions are deleterious, mostly due to their local effects (e.g. that affect gene activity, or that alter chromatin structure) on the host genes. Thus, we wanted to test whether there is a relationship between fixed TE fragments and some key features of their host noncoding sequences, such as length and divergence.

We found that, as expected, there was a significantly positive correlation between length of noncoding sequences and the fraction of sites that are derived from TEs for all intergenic sequences and introns (Spearman's correlation $r = 0.237$, $p < 0.0001$ for the intergenic; Spearman's correlation $r = 0.143$, $p < 0.0001$ for the intronic). Short noncoding sequences (e.g. most introns) are not long enough to contain long canonical TE insertions, especially for LTR and LINE elements/fragments (that usually have very long canonical sequences up to ~10,000bp in length). It has been documented that there is a negative correlation between divergence and length of noncoding sequences in *D. melanogaster* and *D. simulans* (Haddrill et al. 2005; Halligan and Keightley 2006). We

found the same pattern for both intergenic and intronic sequences in *D. melanogaster* and *D. yakuba* (intergenic: Spearman's correlation r = -0.115, $p < 0.0001$; intronic: Spearman's correlation r = -0.255, $p < 0.0001$). We then found that there was a significantly positive correlation between fraction of TE fragments and divergence of orthologous noncoding sequences, although the relationship was not strong (intergenic: Spearman's correlation r = 0.078, $p < 0.0001$; intronic: Spearman's correlation r = 0.021, $p < 0.0001$). It is thus possible that high levels of divergences could simply result from high levels of fractions of TEs in the same non-coding sequences. However, the partial correlation coefficient for divergence versus length of noncoding sequences, controlling for the fraction of TEs, was -0.134 ($p < 0.0001$) for intergenic sequences and -0.261 ($p < 0.0001$) for introns. The partial correlation coefficient for divergence versus fraction of TEs (controlling for length) was 0.121 ($p < 0.0001$) for intergenic sequences and 0.060 ($p < 0.0001$) for introns. The partial correlation coefficient for fraction of TEs versus length (controlling for divergence) was 0.188 ($p < 0.0001$) for intergenic sequences and 0.151 ($p < 0.0001$) for introns. These results suggest that the relationship between length of noncoding sequences and divergence is not a confounding effect of fraction of TEs, despite the positive relationship between divergence and fraction of TEs.

We also wanted to know if any heterogeneity exists in preference of host noncoding sequences between TE classes, by correlating divergence of orthologous TEs and length of noncoding sequences where they reside. We found that there was a significantly negative correlation between divergence of TEs and length of noncoding sequences, and

this correlation was stronger for introns (intergenic: Spearman's correlation $r = -0.046$, $p = 0.0047$; intronic: Spearman's correlation $r = -0.24$, $p < 0.0001$). Low diverged orthologous TEs (e.g., LTR and LINE elements/fragments) tend to reside in relatively longer noncoding sequences than highly diverged TEs (TIR elements/fragments) do. For example, the mean length of intergenic sequences that contain TIR elements/fragments is 31,930bp (95% CI: 30,158bp – 33,704bp), significantly shorter than that of intergenic sequences that contain LTR/LINE elements/fragments, 40,122bp (95% CI: 37,841bp – 42,402bp). This is possibly because canonical lengths of LTR/LINE elements/fragments are, on average, much longer than those of TIR elements/fragments. However, we still found that the number of orthologous TIR elements/fragments was much lower than that of orthologous non-TIR (LTR/LINE) elements/fragments in long noncoding sequences. Our results suggest that many TIR insertions in long noncoding sequences (with higher selective constraints) may have been removed by selection due to their deleterious local effects, and did not become fixed. The presence of relatively larger number of non-TIR elements/fragments in long noncoding sequences possibly implies that some LTR/LINE remnants may have been co-opted for a function as regulatory elements. This may contribute partially to lower divergence in long noncoding sequences.

Representation of TEs in promoter regions, distal promoter regions and proximal 3' regions

We have shown that TE density tends to be higher in positions farther from protein-coding regions (Figure 4.2). We then wanted to investigate the representation of TE fragments in gene regulatory regions (low TE density regions), since it is believed that

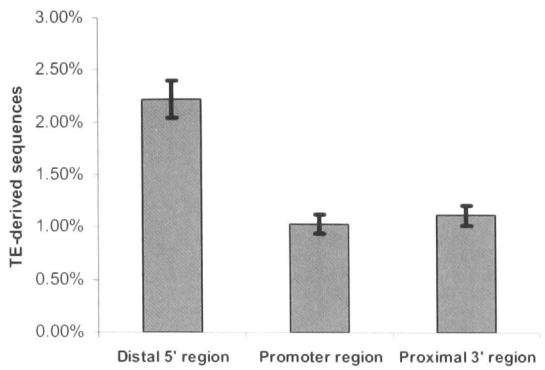


Figure 4.5. - Fraction of TE-derived nucleotide sites in three genomic regions in *Drosophila*, distal 5' regions, promoter regions and proximal 3' regions (500bp in length). 95% CI were shown by bars.

TEs may play important roles in regulating gene expression. Here, we compared TE insertions in promoter regions, as compared to those in adjacent distal 5' regions and proximal 3' regions.

Among 6,763 genes of *D. melanogaster* analyzed, we found that 268 (~4.0%) contain TE-derived sequences in promoter regions, 295 (~4.4%) contain TE fragments in proximal 3' regions and 388 (~5.7%) contain TE fragments in distal 5' regions. These figures are much lower than the fraction of promoter regions that contain TE-derived sequences in human (~25%). Promoter regions also have the lowest total number of nucleotides derived from TE sequences (~1.0%), significantly lower than the percentage of TE-derived nucleotides in distal 5' regions, ~2.2% ($p < 0.0001$, Figure 4.5). This is probably because TE insertions proximal to coding regions are, on average, more likely to be deleterious to the host, and are removed by negative selection. Note that these figures are still much lower than the percentage of nucleotides derived from TEs in

Table 4.2. Representation of TE-derived sequence in promoter regions, distal 5' regions and proximal 3' regions.

Element class	Number of elements/fragments	Length of TE-derived sequences (bp)	Percent of TE-derived sequences
Promoter regions			
LTR	119	12107	0.36%
LINE	34	5116	0.14%
TIR	30	3094	0.09%
INE-1	121	14445	0.43%
Total	304	34762	1.03%
Distal 5' regions			
LTR	145	21660	0.64%
LINE	49	10735	0.32%
TIR	64	14873	0.44%
INE-1	212	27293	0.81%
Total	470	74561	2.21%
Proximal 3'regions			
LTR	121	15416	0.46%
LINE	34	4141	0.12%
TIR	28	4428	0.13%
INE-1	138	13583	0.40%
Total	321	37568	1.11%

LTR: long terminal repeats; LINE: long interspersed nuclear elements; TIR: DNA transposons; INE-1: *Drosophila* interspersed nuclear elements. We separated INE-1 from other DNA transposons since they are the most abundant.

human promoter regions (~8%, Jordan et al. 2003). Note that proximal 3' regions have a very similar fraction of nucleotides derived from TEs (~1.1%) to promoter regions. However, the proportion of TE-derived sequences of those three regions is still much lower than the total proportion of TE-derived sequences of the whole noncoding sequences (~6.0%).

All three regions contain all types of common *Drosophila* TE fragments (Table 4.2), and LTR and INE-1 elements/fragments appear to predominate. It is noteworthy that the fraction of nucleotides derived from LTR is ~2.5 times of that of nucleotides derived

from LINE in promoter and proximal 3' regions (Table 4.2). This is very similar to the ratio of LTR to LINE elements for the whole genome (Figure 4.1). However, since LTR elements or fragments carry more transcription regulating signals than LINE elements/fragments, they are believed to be rarer in gene promoter regions and other regulatory regions (Thornberg et al. 2006; Fablet et al. 2007). This is supported by evidence found in human promoter regions that LTR insertions are much rarer than LINE insertions. In *Drosophila*, however, LTR elements still make up relatively more nucleotides in promoter regions than LINE elements. It is therefore possible that some LTR remnants may have been recruited to be parts of regulatory elements for the host via positive selection in *Drosophila*. Indeed, this is supported by a great body of evidence that LTRs serve as important regulatory elements for many genes and drive genome evolution in humans and *Drosophila* (van de Lagemaat et al. 2003; Kazazian 2004).

However, this does not mean that other classes of TE fragments in promoter regions are not important with respect to donating transcription regulatory signals. On the contrary, using the *Drosophila* transcriptional *cis*-regulatory modules (CDMs) database, REDfly (Gallo et al. 2005), we found that a DNA transposon, *transib4*, overlaps with an enhancer sequence for the salivary gland secretion 3 (*Sgs3*) gene (FBgn0003373), which is associated with structural molecule activity and puparial adhesion (Mourrain et al. 2000). We also found an unknown-typed *D. melanogaster* inverted repeat, *ftz_dm*, overlaps with a zebra element of the fushi tarazu (*ftz*) gene, a *Drosophila* Hox complex (HOM-C) gene, which is associated with transcription factor activity, sequence-specific

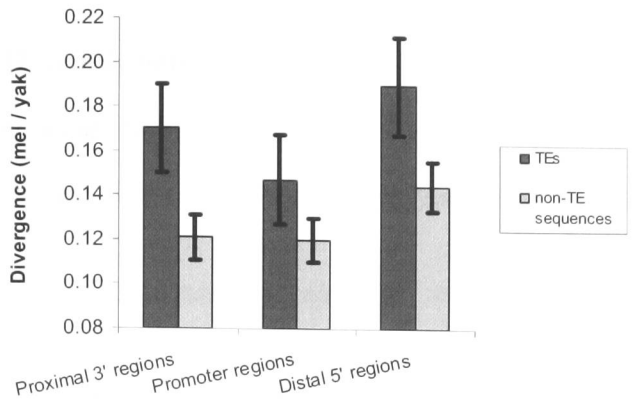
DNA binding, and some other important biological processes (Lohr et al. 2001). These TE elements may have taken part in regulating gene expression for the local genes.

We also found that there was a strong positive relationship between TE density and distance to transcription start/end site in promoter and proximal 3' regions. The extent appears to be stronger when the distance is shortest. This again supports our previous finding that TE insertions are under stronger negative selection (such that they are removed from the population) when they are closest to the transcription start/end site.

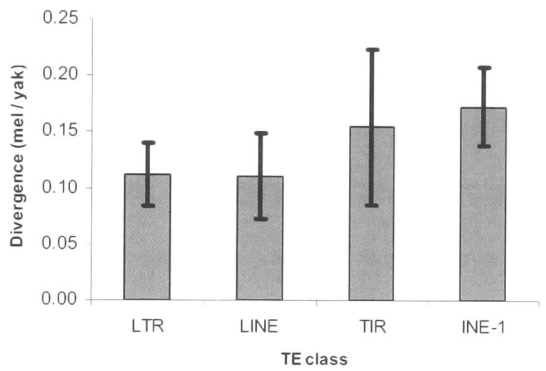
Divergence of TEs among promoter regions, distal 5' regions and proximal 3' regions

In the previous section, we have investigated the differences in representation of major TE classes among promoter regions, distal 5' regions and proximal 3' regions. However, we also wanted to investigate the differences in interspecies divergence for orthologous TE fragments amongst these three genomic regions.

As shown in Figure 4.6.A, orthologous TE fragments in promoter regions have the lowest mean divergence 0.147 [95% CI: 0.126 – 0.167], significantly lower than that of TE fragments in distal 5' regions, 0.189 [95% CI: 0.167 – 0.211, $p < 0.05$]. TE fragments in proximal 3' regions have a similar mean divergence to those in promoter regions. These patterns are still true when the factor of crossing over rate was considered (data not shown). It is possible that either element nesting or biased gene conversion may affect our divergence calculation for the true orthology. But one should note that the regions we chose to compare (promoter regions *vs.* distal promoter regions) are adjacent to each other with only 500bp in length (i.e., they should have similar genomic



(A)



(B)

Figure 4.6. - Mean divergence of orthologous TEs between *D. melanogaster* and *D. yakuba*, for (A) TE-derived sequences and non-TE derived sequences in three genomic regions: proximal 3' regions, promoter regions and distal 5' regions, and (B) TEs only in promoter regions. 95% CI were shown by bars.

background), thus there should be no preference for those events to occur in either of the two regions. It should not affect the relative difference in mean divergence between TE fragments from the two regions greatly, unless genes that contain TE fragments in promoter regions are, on average, more likely to have gene conversion or element nesting than genes that contain TE fragments in distal promoter regions. However,

evidence for this preference remains absent. Our findings at least suggest that, compared to orthologous TE orthologous in distal promoter regions, those in promoter regions (closer to coding sequences) may have a greater potential to donate transcriptional regulatory signals to the host, and have become co-opted in one or both species.

The low interspecies divergence in promoter regions is mainly contributed by retrotransposons, LTR (*roo_I* fragments mostly) and LINE elements, while INE-1 fragments and other DNA transposons showed relatively higher mean divergence (Figure 4.6.B). This difference is significant between INE-1 and LTR fragments ($p < 0.05$). It is noteworthy that orthologous INE-1 fragments in promoter regions also showed lower mean divergence than those in distal 5' regions (data not shown). We also found that mean divergence of TE fragment in promoter regions was lower than that of TE fragments in the whole intergenic regions for each major TE class, although the difference was not significant.

We also calculated mean divergence for nucleotides that are not derived from TEs in all three regions, and compared it with divergence of TE-derived sequences from the same region. Non TE-derived nucleotides always have significantly lower mean divergence than TE-derived nucleotides in all three regions (Figure 4.6.A). This suggests that, even though some TE remnants may be conserved between species and play some regulatory roles, non TE-derived sequences proximal to transcription start/end sites could still contribute the most to regulate gene expression (as promoters or enhancers). It has been suggested that TE-derived sequences serve as a pool of potential regulatory signals that are recruited by the host in response to changes of the genomic environment

(Ludwig et al. 2000; Fablet et al. 2007). However, it is also possible that, since TEs tend to be lineage-specific, TE fragments involved in regulatory functions in *D. melanogaster* may not possess the same functions in *D. yakuba*, and would then evolve relatively faster in *D. yakuba*, or vice versa. It is noteworthy that, consistent with the pattern of TE-derived sequences, non-TE derived sequences in both promoter regions and proximal 3' regions also have significantly lower mean divergence than those in distal 5' regions ($p < 0.05$, Figure 4.6.A). This again suggests that mutations proximal to transcription start/end sites are, on average, deleterious with respect to gene function, and removed by selection.

4.5. Conclusions

The accumulation of genomic sequence data has led to a growing body of research to reveal the roles played by TEs in genome evolution and gene regulation. Here, we first investigated the distribution of three major TE classes (LTR, LINE and TIR elements/fragments) in intergenic, intronic and exonic regions of the *D. melanogaster* euchromatic genome. We found that LTR elements/fragments outnumber the other two TE classes in all regions we studied. Among all LTR elements/fragments, *roo_I* elements/fragments appear to be the most abundant; in particular, ~90% of TEs we recovered from exonic regions are *roo_I* fragments. Genes exhibiting similarities to known TEs in their exons are mostly involved in functions such as transcription factor activity and sequence-specific DNA binding, and/or in biological processes such as cell

differentiation and central nervous system development possibly in response to a stressful environment. We also found that orthologous LTR fragments show the lowest interspecies divergence between *D. melanogaster* and *D. yakuba* compared to orthologous LINE and TIR fragments. In addition, *roo_I* fragments appear to be evolving relatively more slowly than other TE families since the split of *D. melanogaster* and *D. yakuba*, while INE-1 elements/fragments, a family of nonautonomous DNA transposons, are evolving faster than the average of the other element families among TIRs. However, they all show significantly lower mean divergence than synonymous sites, suggesting that TE fragments may be under substantial selective constraints. It is possible that some TE fragments (especially LTRs) are more likely to be co-opted for a function (e.g. gene regulation) by the host, since LTR elements/fragments usually carry more transcriptional regulatory signals than other TE fragments. Thus, they have become more selectively constrained between species. We also showed that orthologous LTR/LINE elements/fragments tend to reside in longer noncoding sequences than orthologous TIR elements/fragments do. Higher selective constraint in long noncoding sequences may be partially due to the presence of more orthologous LTR/LINE elements/fragments.

The most important application of TE fragments being functional is the ability to potentially contribute their regulatory regions to form new host regulatory sequences, particularly when changes are needed to cope with changing genomic environments. We then investigated TE insertions in *Drosophila* promoter regions, compared to distal 5' regions and proximal 3' regions. We found that there are significantly fewer TE

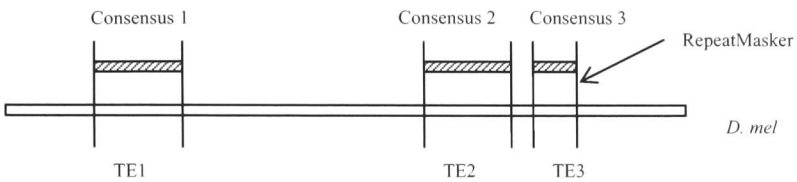
insertions in promoter regions than in the regions compared. TE density increases with the distance to CDS. This is thought to be due to stronger negative selection on deleterious insertions in regions closer to CDS. We also found that TE-derived sequences in promoter regions tend to evolve significantly more slowly than those in distal promoter regions between *D. melanogaster* and *D. yakuba*. TE fragments in promoter regions may have a greater potential to be co-opted if necessary in one or both species. Our findings indicate that TE fragments may have contributed substantially to the gene regulation, and even protein-coding, in *Drosophila*.

4.6. Acknowledgements

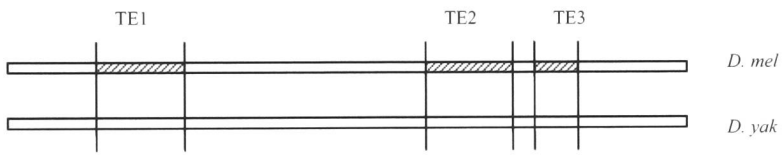
We thank Ensembl and AAA 12 *Drosophila* genomes Project to allow us to use the GenBank data file and the multiple alignments of *Drosophila*. JW was supported by the Dorothy Hodgkin Postgraduate Studentship Award. Funding for DLH was provided by the BBSRC.

4.7. Supplementary Materials

Supplementary Figure 4.1. Precedure of identifying TEs in D. melanogaster and extracting orthologous elements/fragments between D. melanogaster and D. yakuba.



Step 1: Identifying TEs using RepeatMasker. Criteria include
- length of TE > 80bp
- divergence between the consensus sequence and TE < 0.25



Step 2: Identifying orthologous TEs using the AAA 12 *Drosophila* genomes alignments. We carry out
- removing TE alignments whose flanking sequences (100bp in length) contain TE fragments from the same family in *D. mel*, e.g., distance between TE2 and TE3 > 100bp if they are from the same family.
- checking alignments of flanking sequences (50bp in length)
- cleaning up the alignments

Supplementary Figure 4.2. Some examples of orthologous TE alignments.

Example 1: *roo_I* fragment in the third exon of the Atx2 gene.

```
D. melanogaster   CCAGCAGCAACAGCAGCAG---CAGCAGCAGCAGCAGCAGCATCAAGTG
      D. yakuba   CCAGCAGCAGCAGCAGCAGCCACAGCAGCAGCAG---------CAAGTG

D. melanogaster   CAGCAGCAGCAACAGCGAGCGTTGCAGCAATCTGCCTCGCCACCGCAAC
      D. yakuba   CAGCAGCAGCAACAGCGAGCGTTGCAGCAATCTGCCTCGCCACCGCAA-

D. melanogaster   AGCAGCAGCAGCAGCAGCAACAACAGCAGC
      D. yakuba   --------CAACAGCAGCAACAACAGCAAC
```

Example 2: *roo_I* fragment in one of the short exons of the CG31158 gene.

```
D. melanogaster   CAACAGCAAGAGCAGCAATTCCAACAGCAGCAGCAGCAGCTTCACCAG-
      D. yakuba   CAACAGCAAGAGCAGCAATTCCAGCAGCAGCAACAGCAGCTCCACCAG-

D. melanogaster   --CAACATCTGCAGCAACAGCAGCAGCTTCAGCAGCAACATCAGCAGCA
      D. yakuba   --CAACATCTGCAGCAACAGCAGCAACTCCAGCAGCAGCACCAGCAGCA

D. melanogaster   GCAACAACAGCAGCAGC
      D. yakuba   GCAACAACAGCA-----
```

Example 3: *roo_I* fragment in the intergenic region in Chromosome 3R.

```
D. melanogaster   ccgactgcagtagcagcaatggcagcagcg-----ac----gcagcagcg
      D. yakuba   ccgactgcagtagcagcaacagcagcagcgcagcaacagcagcagcagcg

D. melanogaster   acagcaacacgttgcaatgccggtgcagcagcaacatcggcaggagaagc
      D. yakuba   acagcaacacgttgcagtgccggtgcagcagcaacatcggcaggagaagc

D. melanogaster   agcagcagcaacagcaacagcagc
      D. yakuba   agcagcaacagcagcaacggcagc
```

Example 4: INE-1 fragment in the intergenic region in Chromosome 3L.

```
D. melanogaster   gattttaggcaattatatataggaaacacgcat----------------
      D. yakuba   tattttaagcaatcataaataagatatacatatggtccgataaatttgtt

D. melanogaster   -------------tccc-ctgaataactaatttaaatttataaacttcta
      D. yakuba   tagttttagaaactccctttgaataactagcttcaaattggaagtctttg

D. melanogaster   caagaattgcaattagtccgcttagaaaagccgtcttggaaactcgagac
      D. yakuba   gaagaattgcaattagtc-------------------------------a

D. melanogaster   gcctga--------------------aagttcacctgcctggttatcagt
      D. yakuba   gcctgaaagtgcacctgctcttcgagaagttcacctacgtggttatcagt

D. melanogaster   ctgccttgcggcttatcggtcatctaaagttgggctggccaaatgccaca
```

```
      D. yakuba   ctgctttgcggcttatcggtcatctaaagttgggctggccgaatgccaca

D. melanogaster   ggttcgcgggacgctcttaagcggaaacatcccttggcgga---gtgcac
      D. yakuba   ggttcgcgggagactcataagcggaaacatcccttcacagagtggtgcac

D. melanogaster   cacgagtgtgccagaaagtgtgctatggtttttcgcacgctggtggaaat
      D. yakuba   cacgagttcgccggaaagtgtgctatggtttttcgcacgctgttggaaat

D. melanogaster   tcgtggcccggacatcgat
      D. yakuba   tcgtggcccggacatcgat
```

Example 5: DM_CR1A fragment (LINE class) in the intergenic region of Chromosome 2R

```
D. melanogaster   gttgggtacataacggagggagtttttcggcccagttaatgaaagttc--
      D. yakuba   gttgcgtacataacggagggagtttttcggcccagttaatgaaagttcaa

D. melanogaster   ---aaattcgcgtccagaaaaggaaaccggttgattggtt------catc
      D. yakuba   gtgaaattcgcgtttaaaaaaggaaaccggttgattgactcaaccgcatc

D. melanogaster   atcgatttccagatttcgggtccacttatgctcaaggttaaagctgatat
      D. yakuba   atcgatttccagatgtcgggtccacttatgcccaaggttaaa-ctgattt

D. melanogaster   cggcactttgcgtttcttaactgcgtacagttgcacgacttggtctttga
      D. yakuba   aggccatttacatttgttaactgcgtacagttgcacgaaatggtcttaga

D. melanogaster   gccgggaaatctacaggtagatggattttccattgactcagcag------
      D. yakuba   gtagggaattcggcaggtagatgaatttcc-----actctgaagacatac

D. melanogaster   ---------caggtagtcgaaagccagtttgaactggttttgctaggatt
      D. yakuba   atatgtagacaggtagtctgcagccagtttgaactggtttagccagggat

D. melanogaster   ttatatgtgggaaaagactcaa--------------acccagcaggacac
      D. yakuba   gtaattgtgggaaaaggctcactgtcgtgtctgcgtacccggcaggacac

D. melanogaster   taatgttagtttgattggtttattgtttgtcgttcac
      D. yakuba   aaatgtcagttggactgttttattatttgtcgtttac
```

Chapter 5.

Discussion and Conclusions

In eukaryotes, a large percentage of the total genome size of many organisms is comprised of noncoding DNA, known as the "C-value enigma" (Gregory 2001; Gregory 2005, p.3-87). The human genome, for example, comprises only about 1.5% protein-coding DNA sequences, with the other 98.5% being various types of noncoding DNA ((International Human Genome Sequencing Consortium 2001). Among these noncoding DNA sequences, ~50% are constituted by transposable elements (TEs). Comparative sequence analysis on a genomic scale has opened the door to the systematic analysis of the relationship between noncoding DNA and biological/evolutional meaning. Estimating the fraction of noncoding DNA that is functionally important will help solve the C-value paradox/enigma, although it has been long thought that most noncoding DNA in multicellular eukaryotic genomes is unconstrained.

Interspecies sequence comparisons of noncoding regions reveal conserved features, many of which are likely to be *cis*-regulatory elements (Hardison 2000; Ludwig 2002). The recent genome-wide analysis of divergence and selective constraints in the noncoding sequences of yeast, *Drosophila* and mammals has revealed some unexpected evolutionary features of non-protein-coding DNA sequences. In yeast, regions of high conservation account for ~34% of the total 5′ upstream sequences analyzed (Chin et al. 2005). In *Drosophila*, Halligan and Keightley (2006) have estimated that >50% of point substitutions in intergenic and intronic sequences are removed by negative selection.

Furthermore, 5′ and 3′ flanking sequences up to 5kb away from coding sequence boundaries still appear to be under substantial selective constraints (~0.60) (Halligan and Keightley 2006). On the contrary, in murids and hominids, functional non-coding elements tend to be clustered mostly within 2kb surrounding protein-coding sequences (Keightley et al. 2005). It is likely that many/some TEs in these regions are selectively constrained, serving as *cis*-regulatory elements.

During the study of functional sequence evolution, the process of sequence alignment is often unmentioned, since it is a particularly difficult task when noncoding sequences are involved. Different alignment schemes can produce very different alignments, and produce very different results concerning the rate of substitutions. Comparison of a large amount of non-homologous nucleotide bases will make any subsequent analysis meaningless. Although numerous alignment tools have been developed to align genomic sequences, many of them appear to be inaccurate for aligning noncoding sequences, due to the relatively higher frequencies of indels in noncoding sequences than in protein-coding sequences. Pollard et al. (2004) have benchmarked eight pairwise alignment tools for aligning simulated *Drosophila* noncoding sequences, and found that AVID (Bray et al. 2003) and CLUSTALW (Thompson et al. 1994) have higher sensitivity over entire noncoding sequences as well as in constrained sequences, compared to the others. A stochastic model-based alignment method, MCALIGN (Keightley and Johnson 2004), appears to outperform both AVID (a heuristic method) and CLUSTALW (a score-based heuristic scheme), since it incorporates empirical information for distribution of indel

lengths and rates of indels in relation to point substitutions, making alignments more biologically meaningful.

In this project, an accurate global pairwise alignment method, MCALIGN2, has been developed for noncoding DNA sequence alignment. It is based on explicit models of indel evolution. The improvement of this method in relation to MCALIGN is that it employs a time-continuous pair HMM of seven states, offering a well-performed approximation to the "real" evolutionary process. Moreover, the optimizer in MCALIGN2 is guaranteed to find the optimal alignment for a given divergence time based on dynamic programming, and then to approach to the most probable alignment faster than the Monte Carlo method used in MCALIGN. This gives MCALIGN2 advantages over MCALIGN when aligning relatively diverged/long sequences in terms of accuracy/speed.

The common feature of the two methods is that, instead of specifying match/mismatch/gap penalty parameters and scoring matrices for all unknown variables (e.g. in CLUSTALW), both rest on Bayesian statistics in which all the variables (observed data and the unknowns) in an inference problem are treated as random variables. There are two unknown variables, divergence time t and alignment a. We focus on alignment a, treating time t as a nuisance parameter, and obtain $P(\textit{one unknown } `a' \mid \textit{data})$ by integrating the posterior distribution as shown in Equation 2.2. Following this step, we can evaluate the appropriateness of the model and suggest improvements for alignment a based on $P(a \mid \textit{data})$. Besides the relaxation of the traditional fixed parameter settings, the other advantageous feature of Bayesian inference is that it fully

accounts for the uncertainty of all unknowns in the posterior distribution (Liu and Lawrence 1999). It uses the posterior probability as the guiding principle to manipulate data and information. This probability-based model has been proved to be coherent and efficient for quantifying objective and subjective uncertainties, and has been accepted as an appropriate and competent method in almost all information-based technologies (Zhu et al. 1998; Liu and Lawrence 1999), such as economic evaluation of projects (Allen 1991) and quantifying the climate system (Forest et al. 2002).

It has been shown that MCALIGN2 outperforms other available sequence alignment methods, using both simulated and real noncoding sequence data. More complicated/realistic models have been developed recently (e.g., a "long indel" model), but they tend to be extensively computationally expensive, making aligning long noncoding DNA sequences (up to several mega base pairs) realistically impossible. MCALIGN2 has been used to analyze genomic sequences in *Drosophila* (Halligan and Keightley 2006; Haddrill et al. 2007), rice (Guo et al. 2007) and *Arabidopsis* (DeRose-Wilson and Gaut 2007), and it has been downloaded by more than 500 users.

Comparative genomic analysis has offered a great opportunity to unravel the functional importance of noncoding DNA sequences, of which transposable elements are one of the major components in many eukaryotic organisms. It has long been thought that TEs are just genomic selfish elements, adding new copies of themselves into the genome regardless of the consequences (Bushman 2004). This unique feature of TEs may, however, help in normal gene regulation. We investigated the distribution and patterns of evolution (e.g., interspecies divergence) of TE-derived sequences from three major

classes (LTR, non-LTR retrotransposons and DNA transposons) in the *Drosophila* euchromatic genome, using a gene-centric approach by comparing TE fragments among intergenic, intronic and exonic regions. Our main findings are listed and discussed as follow:

(1) As shown in previous studies, TEs are not randomly distributed within the genome. LTR elements outnumber LINE and DNA (also called TIR) elements in all intergenic, intronic and exonic regions. Among LTR retrotransposon, *Pao* family elements, *roo_I*, appear to be the most abundant, in particular in exonic regions (that is ~90% of TEs in these regions are *roo_I*). We also found that LTR/LINE fragments tend to reside within relatively longer intergenic and intronic sequences than TIR fragments do.

(2) Orthologous LTR/LINE fragments show a significantly lower mean interspecies divergence than orthologous TIR fragments between *D. melanogaster* and *D. yakuba*, and of these, *roo_I* fragments appear to be evolving the most slowly. We also showed that the mean divergence of all TE fragments is significantly lower than that of synonymous sites, but higher than that of non-synonymous sites in *D. melanogaster* and *D. yakuba*. This suggests some TE fragments are selectively constrained, and may have been co-opted by the host for a function.

(3) ~1.0% of nucleotides in promoter regions of *D. melanogaster* are derived from TE sequences. TE density is relatively higher in regions proximal to coding

sequences than in distal regions. TEs proximal to CDS also show relatively lower mean divergence than those farther from CDS.

(4) INE-1 elements are among the fastest evolving TEs in *Drosophila*. They appear to be evolving the fastest between *D. melanogaster* and *D. simulans*, faster than all four-fold sites. However, they appear to be evolving significantly more slowly than synonymous sites between *D. melanogaster* and *D. yakuba*. This may be because INE-1 elements have recently become active in the lineage of *D. yakuba*.

(5) Interspecies divergence of orthologous TE remnants is not simply due to genetic drift of neutral mutations that occurred after the divergence. Divergence could also be affected by ancestral polymorphisms, and some form of natural selection. Divergence of some fast evolving TEs and other fast evolving sites (e.g. sites within short introns and four-fold degenerate sites) is strongly influenced by the recombination environment in which they are located. The positive correlation between divergence and crossing over frequency suggests that differences in the level of the ancestral polymorphism could be the major force of determining the interspecies divergence, whereas the negative correlation between divergence and frequency of crossing over could be an indication of some form of natural selection operating on sites.

Our findings suggest that the evolution of transposable elements could be a very complicated process that is not simply determined by any single evolutionary force.

Moreover, since TEs are diversely categorized due to the different transpositional mechanisms, evolutionary patterns could be quite different among those TE classes. This is supported by our findings that the distribution and interspecies divergence are quite different among TEs from the three major classes. Furthermore, LTR elements carry more transcriptional regulatory signals than LINE/DNA elements, thus, they may have more potential to be co-opted by the host when domestication is needed. This is supported by the observation that orthologous LTR elements show the lowest mean divergence, resting on the hypothesis that functionally important nucleotides tend to be conserved between species and show relatively lower interspecies divergence compared to relatively unconstrained sites. However, the conservation of the same TEs could differ among close species, since TEs tend to be lineage-specific. This is shown by the difference in divergence between orthologous INE-1 elements of *D. melanogaster* and *D. simulans* and those of *D. melanogaster* and *D. yakuba*, compared to the same class of sites. Overall, our findings suggest that TEs may have contributed substantially to the host genome evolution, by donating their own regulatory regions to form host regulatory sequences, or even coding for host proteins. This co-option process is possibly stimulated by the changing environmental conditions (e.g. caused by migration)

In the future, it will be interesting to develop a multiple alignment algorithm, based on the pairwise alignment algorithm and explicit models of indel evolution. Bayesian coestimation of phylogeny and sequence alignment has been proved to be the most appropriate approach for multiple alignments (Sankoff et al. 1973; Sankoff and Cedergren 1983; Lunter et al. 2005), because phylogeny and alignment are

interdependent, and coestimation accounts for all the variables. However, this method tends to be very computationally expensive, with the complexity increasing dramatically with numbers and lengths of sequences. The alternative is to use a heuristic method, the progressive alignment strategy (Feng and Doolittle 1987; Barton and Sternberg 1987; Higgins and Sharp 1989), by aligning the most similar pairs of sequences first and aligning subsequent sequences according to a "guide tree". The multiple alignment algorithms will help carry out more comprehensive analysis on the evolution of noncoding DNA sequences, e.g., to search for conservative blocks within noncoding sequences among several related species and to further infer the ancestral structure.

Our findings support previous intuitions that noncoding DNA sequences have played very important roles in gene regulation, in particular for TEs, and some of them may even contribute their coding potential to the host. Transposable elements have shaped both host genes and the host genome, and they have become a useful and powerful tool for understanding more about gene regulation/functions and genome evolution. One should note that to align homologous/orthologous noncoding sequences carefully and correctly is crucial to reveal the true evolutionary signature hidden in the vast amount of noncoding DNA. Thus, more attention should be paid to choose proper alignment methods and/or to design better evolutionary models for indels and point substitutions for particular sequences under study. Any subsequent analysis will then reveal the "true" evolutionary patterns based on the quality alignments. More and more attention and effort have been paid to study noncoding DNA sequences since the emergence and availability of whole genome sequences of many species. Hopefully, more surprising

and currently unknown evolutionary features of noncoding DNA will be discovered in the near future.

References

Adams, M.D. et al. 2000. The genome sequence of Drosophila melanogaster. *Science* **287**: 2185-2195.

Akashi, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. *Genetics* **139**: 1067-1076.

Almeida, L.M., Silva, I.T., Silva, W.A., Castro, J.P., Riggs, P.K., Carareto, C.M. and Amaral, M.E.J. 2007. The contribution of transposable elements to *Bos taurus* gene structure. *Gene* **390**: 180-189.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403-410.

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

Aquado, C.F., Begun, D.J. and Kindahl, E.C. 1994. Selection, recombination, and DNA polymorphism in Drosophila. *In Non-neutral Evolution: Theories and Molecular Data* (ed. B. Golding), pp. 46-56. London: Chapman and Hall.

Allen, D.H. 1991. *Economic evaluation of projects: a guide.* Institution of chemical engineers (IChemE). Capital investments. ISBN: 085295266X.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., et al. 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science* **306**: 79-86.

Arnault, C. and Dufournel, I. 1994. Genome and stresses: reactions against aggressions, behavior of transposable elements. *Genetica* **93**: 149-160.

Ashburner, M. and Rubin, G.M. *Drosophila melanogaster*: a case study of a model genomic sequence and its consequences. *Genome Res.* **15**: 1661-1667.

Bannert, N., and Kurth, R. 2004. Retroelements and the human genome: new perspectives on an old relation. *Proc. Natl. Acad. Sci. U. S. A.* **5**: 14572-14579.

Barinaga, M. 1997. Cells count proteins to keep their telomeres in line. *Science* **275**: 928.

Bartolomé, C., Maside, X. and Charlesworth, B. 2002. On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster. *Mol Biol Evol.* **19**: 926-937.

Barton, G.J. and Sternberg, M.J.E. 1987. A strategy for the rapid multiple alignment of protein sequences. *Journal of Molecular Biology* **198**: 327-337.

Beaton, M.J. and Cavalier-Smith, T. 1999. Eukaryotic non-coding DNA is functional: evidence from the differential scaling of cryptomonal genomes. *Proc. R. Soc. Lond. B.* **266**:2053-2059.

Begun, D.J. and Aquadro, C.F. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* **356**: 519-520.

Bejerano, G., Lowe C. B., Ahituv N., King B., Siepel, A. *et al.*, 2006 A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**: 87-90.

Berg, D.E. and Howe, M.M. 1989. *Mobile DNA.* ASM Press.

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

Bergman, C.M. and Bensasson, D. 2007. Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in *Drosophila melanogaster. Proc Natl Acad Sci U S A* **104**(27): 11340-45.

Bermano, G., Arthur, J.R. and Hesketh, J.E. 1996. Role of the 3' untranslated region in the regulation of cytosolic glutathione peroxidase and phospholipid-hydroperoxide glutathione peroxidase gene expression by selenium supply. *Biochem J* **320**: 891-895.

Betancourt, A.J., and Presgraves, D.C. 2002. Linkage limits the power of natural selection in Drosophila. *Proc Natl Acad Sci U S A* **99**: 13616-13620.

Biémont, C., Tsitrone, A., Vieira, C., and Hoogland, C. 1997. Transposable element distribution in *Drosophila. Genetics* **147**: 1997-1999.

Biémont, C. and Cizeron, G. 1999. Distribution of transposable elements in *Drosophila* species. *Genetica* **105**: 43-62.

Biémont, C. and Vieira, C. 2006. Genetics: Junk DNA as an evolutionary force. *Nature* **443**: 521-524.

Borts, R.H. and Haber, J.E. 1987. Meiotic recombination in yeast: Alteration by multiple heterozygosities. *Science* **237**: 1459-1465.

Bouhassira, E.E., Kielman, M.F., Gilman, J., Fabry, M.F., Suzuka, S., Leone, O., Gikas, E., Bernini, L.F., Nagel, R.L. 1997. Properties of the mouse alpha-globin HS-26: relationship to HS-40, the major enhancer of human alpha-globin gene expression. *Am J Hematol* **54**: 30-39.

Bray, N., Dubchak, I. and Pachter, L. 2003. AVID: A global alignment program. *Genome Res.* **13**:97-102.

Bray, N., and Pachter, L. 2004. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res* **14**: 693-699.

Brett, D., Pospisil, H., Valcárcel, J., Reich, J. and Bork, P. 2001. Alternative splicing and genome complexity. *Nature Genetics* **30**: 29-30.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A. and Batzoglou, S. 2003. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**: 721-731.

Bushman, F. 2004. Gene regulation: selfish elements make a mark. *Nature* **429**: 253-255.

Capy, P., Gasperi, G., Biémont, C. and Bazin, C. 2000. Stress and transposable elements: co-evolution or useful parasites? *Heredity* **85**: 101-6.

Castillo-Davis, C.I., and Hartl, D.L. 2003. GeneMerge – post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics* **19**(7):891-892.

Castillo-Davis, C.I. 2005. The evolution of noncoding DNA: how much junk, how much func? *Trends Genet.* **21**: 533-536.

Cavalier-Smith, T. 1985. *The Evolution of Genome Size.* John Wiley, New York.

Celniker, S.E. et al. 2002. Finishing a whole-genome shotgun: Release 3 of the Drosophila melanogaster euchromatic genome sequence. *Genome Biol.* **3**: research0079.1-0079.14.

Celniker, S.E. and Rubin, G.M. The *Drosophila melanogaster* genome. *Annu. Rev. Genomics Hum. Genet.* **4**: 89-117.

Charlesworth, B. and Charlesworth, D. 1983. The population dynamics of transposable elements. *Genet. Res.* 42: 1-27.

Charlesworth, B. and Lapid, A. 1989. A study of ten transposable elements on X chromosomes from a population of Drosophila melanogaster. *Genet. Res.* **54**: 113-125.

Charlesworth, B., and Langley, C. H. 1989. The population genetics of Drosophila transposable elements. *Annu Rev Genet* **23**: 251-287.

Charlesworth, B., Lapid, A. and Canada, D. 1992. The distribution of transposable elements within and between chromosomes in a population of Drosophila melanogaster. I. Element frequencies and distribution. *Genet. Res.* **60**: 103-114.

Charlesworth, B., 1996. Background selection and patterns of genetic diversity in Drosophila melanogaster. *Genet. Res.* **68**: 131-149.

Charlesworth, B., Langley, C.H. and Sniegowski, P.D. 1997. Transposable element distributions in *Drosophila. Genetics* **147**: 1993-1995.

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference. *J. R. Statist. Soc. A.* **158**: 419-466.

Chin, C.-S. Chuang, J.H. and Li, H. 2005. Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. *Genome Res.* **15**: 205-213.

Cost, G.J., Feng, Q., Jacquier, A. and Boeke, J.D. 2002. Human L1 element target-primed reverse transcription *in vitro*. *EMBO J.* **21**: 5899-5910.

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. 1978. A model of evolutionary change in proteins. In Atlas of Protein Sequence and Structure (Dayhoff, M. O., ed.), volume **5**, Suppl.3, pp. 345-352. National Biomedical Research Foundation, Silver Spring, Washington D.C.

Deaconescu, A.M., Chambers, A.L., Smith A.J., Nickels B.E., Hochschild, A., *et al.*, 2006. Structural basis for bacterial transcription-coupled DNA repair. *Cell* **124**: 507-520.

DeBarry, J.D., Ganko, E. and McDonald, J.F. 2005. The contribution of LTR retrotransposon sequences to gene evolution in *Mus musculus*. *Mol. Biol. Evol.* **23**: 479-481.

Deininger, P.L., and Batzer, M.A. 2002. Mammalian retroelements. *Genome Res* **12**: 1455-1465.

Deininger, P.L., Moran, J.V., Batzer, M.A. and Kazazian, Jr., H.H. 2003. Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* **13**: 651-658.

DeRose-Wilson, L.J. and Gaut, B.S. 2007. Transcription-related mutations and GC content drive variation in nucleotide substitution rates across the genomes of *Arabidopsis thaliana* and *Arabidopsis lyrata*. *BMC Evol Biol.* **7**: 66.

Dermitzakis, E.T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B.J., et al. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420**: 578-582.

Dermitzakis, E.T., Reymond, A., Scamuffa, N., Ucla, C., Kirkness, E., Rossier, C. and Antonarakis, S.E. 2003. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* **302**: 1033-1035.

Dermitzakis, E.T., Kirkne, E., Schwarz, S., Birney, E., Reymond, A. and Antonarakis, S.E. 2004. Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraints is independent of their genic environment. *Genome Res.* **14**: 852-859.

Dimitri, P. 1997. Constitutive heterochromatin and transposable elements in *Drosophila melanogaster*. *Genetica* **100**: 85-93.

Dimitri, P. and Junakovic, N. 1999. Revising the selfish DNA hypothesis: new evidence on accumulation of transposable element in heterochromatin. *Trends Genet* **15**: 123-124.

Drake, A., Bird, C., Nemesh, J., Thomas, D.J., Newton-Cheh, C., et al. 2005. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature Genetics* **38**: 223-227.

Drosophila 12 Genomes Consortium 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**: 203-218.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, chapters 2, 3 and 4. Cambridge University Press, Cambridge, UK.

Duret, L., Marais, G. and Biémont, C. 2000. Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. *Genetics* **156**(4): 1661-9.

Eddy, S.R. 2002. Computational genomics of noncoding RNA genes. *Cell* **109**: 137-140.

Eickbush, T.H. and Malik, H.S. 2001. Evolution of retrotransposons, in Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M. Eds. *Mobile DNA II*, Chap. 47. American Society for Microbiology, Washington, DC.

Eickbush, T.H. and Furano, A.V. 2002. Fruit flies and humans respond differently to retrotransposons. *Curr Opin Genet Dev.* **12**: 669-674.

Ewens, W.J. and Grant, G.R. 2001. *Statistical Methods in Bioinformatics*. Springer-Verlag, New York.

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

Felsenstein, J., 1974 The evolutionary advantage of recombination. *Genetics* **78**: 737-756.

Felsenstein, J. 2004. *Inferring Phylogenies*, chapter 13. Sinauer Associates, Sunderland, MA.

Feng, D.-F. and Doolittle, R.F. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution* **25**: 351-360.

Flavell, R.B. 1986. Repetitive DNA and chromosome evolution in plants. *Philos. Trans. R. Soc. Lond. Biol. Sci.***321**: 227-242.

Flavell, R.B, Bennett, M.D, Smith, J.B., and Smith, D.B. 1994. Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem. Genet.* **12**: 257-269.

Forest, C.E., Stone, P.H., Sokolov, A.P., Allen, M.R. and Webster, M.D. 2002. Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science* **295**: 113-117.

Frazer, K.A., Sheehan, J.B., Stokowski, R.P., Chen, X.Y., Hosseini, R., Cheng, J.F., Fodor, S.P.A., Cox, D.R. and Patil, N. 2001. Evolutionary conserved sequences on human Chromosome 21. *Genome Res.* **11**: 1651-1659.

Gaffney, D.J., and Keightley, P.D. 2006. Genomic selective constraints in murid noncoding DNA. *PLoS Genet* **2**: e204.

Gallo, S.M., Li, L., Hu, Z., and Halfon, M.S. 2005. REDfly: a regulatory element database for *Drosophila*. Bioinformatics 22: 381-383.

Ganko, E.W., Greene, C.S., Lewis, J.A., Bhattacharjee, V. and McDonald, J.F. 2006. LTR retrotransposon-gene associations in *Drosophila melanogaster*. *J. Mol. Evol.* **62**: 111-120.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. 2003. *Bayesian Data Analysis*, chapter1 and 12. Chapman and Hall/CRC Press, New York.

Gregory, T.R. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biological Reviews 76: 65-101.

Gregory, T.R. 2005. *The Evolution of The Genome*. Elsevier, ISBN 0-12-301463-8.

Griffiths, A.J.F., Gelbart, W.M., Miller, J.H. and Lewontin, R.C. 1999. The molecular basis of mutation in *Modern Genetic Analysis*, Chapter 7, published by W.H. Freeman and Company ISBN 0-7167-3597-0.

Guo, X., Wang, Y., Keightley, P.D. and Fan, L. 2007. Patterns of selective constraints in noncoding DNA of rice. *BMC Evolutionary Biology* **7**: 208.

Haddrill, P.R., Charlesworth, B., Halligan, D.L. and Andolfatto, P. 2005. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biology* **6**: R67.

Haddrill, P.R., Halligan, D.L., Tomaras, D. and Charlesworth, B. 2007. Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over. *Genome Biology* **8**: R18.

Halligan, D.L., Eyre-Walker, A., Andolfatto, P. and Keightley, P.D. 2004 Patterns of evolutionary constraints in intronic and intergenic DNA of Drosophila. *Genome Res.* **14**: 273-279.

Halligan, D.L., and Keightley, P.D. 2006. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Res* **16**: 875-884.

Handen JS, Rosenberg HF. 1997. Intronic Enhancer Activity of the Eosinophil-derived Neurotoxin (RNS2) and Eosinophil Cationic Protein (RNS3) Genes Is Mediated by an NFAT-1 Consensus Binding Sequence. *J Biol. Chem.* **272**: 1665-1669.

Hardison, R.C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet* **17**: 637-645.

Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., *et al.*, 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* **13**: 13-26.

Harrington L, McPhail T, Mar V, Zhou W, Oulton R, Bass MB, Arruda I, Robinson MO. 1997. A mammalian telomerase-associated protein. *Science* **275**: 973-977.

Hellmann, I., Ebersberger, I., Ptak, S.E., Pääbo, S. and Przeworski, M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* **72**: 1527–1535.

Hey, J., and Kliman, R.M. 2002. Interactions between natural selection, recombination and gene density in the genes of Drosophila. *Genetics* **160**: 595-608.

Higgins, D.G. and Sharp, P.M. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. *Computer Applications in the Bioscience* **5**: 151-153.

Hill, W.G., and Robertson, A. 1966. The effect of linkage on limits to artificial selection. *Genet Res* **8**: 269-294.

Holmes, I. and Bruno, W.J. 2001. Evolutionary HMMs: A Bayesian approach to multiple alignment. *Bioinformatics* **17**: 803-810.

Holt, R.A. et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**: 129-149.

Hudson, R. R. 1991. Gene genealogies and the coalescent process, pp. 1-44 in *Oxford Surveys in Evolutionary Biology*, edited by D. J. Futuyama and J. Antonovics. Oxford University Press, Oxford.

Hüttenhofer, A., et al. 2002. RNomics: identification and function of small, non-messenger RNAs. *Curr. Opin. Chem. Biol.* **6**: 835-843.

Hüttenhofer, A., Schattner, P. and Polacek, N. 2005. Non-coding RNAs: hope or hype. *Trends Genet.* **21**: 289-297.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

International Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.

Jakubczak, J.L., Xiong, Y. and Eickbush, T.H. 1990. Type I (R1) and type II (R2) ribosomal DNA insertions of Drosophila melanogaster are retrotransposable elements closely related to those of Bombyx mori. *J Mol Biol* **212**: 37-52.

Jensen, M.A., Charlesworth, B. and Kreitman, M. 2002. Patterns of genetic variation at a chromosome 4 locus of Drosophila melanogaster and D. simulans. *Genetics* **160**: 493-507.

Jordan, I.K., Rogozin, I.B., Glazko, G.V. and Koonin, E.V. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* **19**: 68-72.

Jukes, T.H. and Cantor, C.R. 1969. Evolution of protein molecules. In *Mammalian protein metabolism* (ed. H.N. Munro), pp. 21-123. Academic Press, New York.

Kaminker, J.S., Bergman, C.M, et al. 2002. The transposable elements of the Drosophila melanogaster euchromatin: a genomic perspective. *Genome Biol.* **3**: 0084.1-0084.2.

Kapitonov, V.V., and Jurka, J. 1999. DNAREP1_DM. Repbase update Release 3.4 (www.girinst.org/Repbase_Updata.html), pp.

Kapitonov, V.V., and Jurka, J. 2003. Molecular paleontology of transposable elements in the Drosophila melanogaster genome. *Proc Natl Acad Sci U S A* **100**: 6569-6574.

Kaplan, N.L. and Brookfield, J.F.Y. 1983. Transposable elements in Mendelian populations. III. Statistical results. *Genetics* **104**: 485-495.

Kaplan, N.L., Hudson, R.R., and Langley, C.H. 1989. The "hitch-hiking effect" revisited. *Genetics* **123**(4): 887-899.

Karlin, S. and Altschul, S.F. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. U.S.A.* **90**: 5873-5877.

Kazazian, H.H. 2004 Mobile Elements: Drivers of Genome Evolution. *Science* **303**(5664): 1626–1632.

Keightley, P.D. and Gaffney, D.J. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc. Natl. Acad. Sci.* **100**: 13402-13406.

Keightley, P.D. and Johnson, T. 2004. MCALIGN: Stochastic alignment of noncoding DNA sequences based on an evolutionary model of sequence evolution. *Genome Res.* **14**: 442-450.

Keightley, P.D., Lercher, M.J. and Eyer-Walker, A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* **3**: e42.

Keplinger BL, Rabetoy AL, Cavener DR. 1996. A somatic reproductive organ enhancer complex activates expression in both the developing and the mature *Drosophila* reproductive tract. *Dev Biol* **180**: 311-323.

Kidwell, M.G. and Lisch, D.R. 1997. Transposable elements as sources of variation in animals and plants. *Proc Natl Acad Sci U S A* **94**: 7704-7711.

Kidwell, M.G. and Lisch, D.R. 2000. Transposable elements and host genome evolution. *Trends Ecol Evol.* **15**: 95-99.

Kidwell, M.G. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**: 49-63.

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* **217**: 624-626.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16:111–120.

Kirby, D.A., Muse, S.V. and Stephan, W. 1995. Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc Natl Acad Sci U S A* **92**: 9047-9051.

Kliman, R.M., and Hey, J. 1993. Reduced natural selection associated with low recombination in Drosophila melanogaster. *Mol Biol Evol* **10**: 1239-1258.

Knudsen, B. and Miyamoto, M.M. 2003. Sequence alignments and pair hidden markov models using evolutionary history. *J.Mol.Biol.* **333**: 453-460.

Kohler J, Schafer-Preuss S, Buttgereit D. 1996. Related enhancers in the *intron* of the beta1 tubulin gene of *Drosophila melanogaster* are essential for maternal and CNS-specific expression during embryogenesis. *Nucleic Acids Res* **24**: 2543-2550.

Kreahling, J. and Graveley, B.R. 2004. The origin and implications of Aluternative splicing. *Trends Genet.* **20**: 1-4.

Labrador, M. and Corces, V.G. 1997. Transposable element-host interaction: Regulation of insertion and excision. *Annu Rev Genet.* **31**: 381-404.

Lanave, C., Preparata, G., Saccone, C. and Serio, G. 1984. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution* **20**:86-93.

Langley, C.H., Montgomery, E.A., Hudson, R., Kaplan, N. and Charlesworth, B. 1988. On the role of unequal exchange in the containment of transposable element copy number. *Genet. Res.* **52**: 223-236.

LePage, D.F., Church, D.M., Millie, E., Hassold, T.J. and Conlon, R.A. 2000. Rapid generation of nested chromosomal deletions on mouse chromosome 2. *Proc Natl Acad Sci U S A* **97**: 10471-10476.

Lercher, M.J., and Hurst, L.D. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* **18**: 337-340.

Li, W.-H. and Graur, D. 1991. Fundamentals of Molecular Evolution. Sinauer Associations, Sunderland, MA.

Li, W.H. 1997. *Molecular Evolution.* Sinauer Associates, Sunderland, MA.

Liu, G., Zhao, S., Bailey, J.A., Sahinalp, S.C., Alkan, C. et al. 2003. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* **13**: 358-368.

Liu, J.S. and Lawrence, C.E. 1999. Bayesian inference on biopolymer models. *Bioinformatics* **15**: 38-52.

Lohr, U., Yussa, M., and Pick, L. (2001). *Drosophila* fushi tarazu: a gene on the border of homeotic function. *Curr. Biol.* **11**: 1403--1412.

Luan, D.D., Korman, M.H., Jakubczak, J.L. and Eickbush, T.H. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. *Cell* **72**: 595-605.

Ludwig, M.Z., Bergman, C., Patel, N.H., and Kreitman, M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403:** 564-567.

Ludwig, M.Z. 2002. Functional evolution of noncoding DNA. *Curr Opin Genet Dev.* **12**: 632-639.

Lunter, G.A., Drummond, A.J., Miklós, I. and Hein, J. 2004. Statistical Alignment: Recent Progress, New Applications, and Challenges, in: Rasmus Nielsen (ed.), "*Statistical methods in Molecular Evolution*", Springer Verlag's Series in Statistics in Health and Medicine.

Lunter, G., Miklos, I., Drummond, A., Jensen, J.L. and Hein, J. 2005. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* **6**: 83.

Marais, G., Domazet-Loso, T., Tautz, D. and Charlesworth, B. 2004. Correlated evolution of synonymous and nonsynonymous sites in Drosophila. *J Mol Evol* **59**: 771-779.

Mattick, J.S. 2004. RNA regulation: a new genetics? *Nat. Rev. Genet.* **5**: 316-323.

Mariño-Ramírez, L., Lewis, K.C., Landsman, D., and Jordan, I.K. 2005. Transposable elements donate lineage-specific regulatory sequences to host genoms. *Cytogenet Genome Res.* **110**: 333-341.

Maynard-Smith, J. and Haigh, J. 1974. The hitch-hiking effect of a favorable gene. *Genet Res* **23**: 23–35.

Maside, X., Assimacopoulos, S. and Charlesworth, B. 2000. Rates of movement of transposable elements on the second chromosome of *Drosophila melanogaster*. *Genet Res* **75**: 275-84.

Maside, X., Bartolome, C., Assimacopoulos, S. and Charlesworth, B. 2001. Rates of movement and distribution of transposable elements in *Drosophila melanogaster*: in situ hybridization vs Southern blotting data. *Genet Res.* **78**: 121-36.

McClintock, B. 1944. The relationship of homozygous deficiencies to mutations and allelic series in maize. *Genetics* **29**: 478-502.

McClintock, B. 1950. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci USA*. **36**: 344-355.

McClintock, B. 1951. Chromosome organization and genic expression. *Cold Spring Harb Symp Quant Biol.* **16**: 13-47.

McDonald, J.F., 1993. Evolution and consequences of transposable elements. *Curr Opin Genet Dev* **3**: 855-864.

McGowan, K.M., Police, S., Winslow, J.B. and Pekala, P.H. 1997. Tumor necrosis factor-alpha regulation of glucose transporter (GLUT1) mRNA turnover. Contribution of the 3'-untranslated region of the GLUT1 message. *J Biol Chem* **272**: 1331-1337.

McVean, G.A., and Vieira, J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. *Genetics* **157**: 245-257.

Miklos, I. and Toroczkai, Z. 2001. *An improved model for statistical alignment.* Pp. 1-10. in First workshop on algorithms in bioinformatics. Springer-Verlag, Berlin, Heidelberg.

Miklos, I., Lunter, G.A. and Holmes, I. 2004. A "long indel" model for evolutionary sequence alignment. *Mol. Biol. Evol.* **21**(3): 529-540.

Miller, W. and Myers, E.W. 1988. Sequence comparison with concave weighting functions. *Bulletin of Mathematical Biology* 50: 97-120.

Misra, S. et al. 2002. Annotation of the *Drosophila melanogaster* genome: a systematic review. *Genome Biol.* **3**: RESEARCH0083.

Montgomery, E.A., Huang, S.M., Langley, C.H. and Judd, B.H. 1991. Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*: genome structure and evolution. *Genetics* **129**: 1085-98.

Morgenstern, B. 1999. DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**: 211-218.

Moriyama, E.N. and Powell, J.R. 1996. Intraspecific nuclear DNA variation in Drosophila. *Mol. Biol. Evol.* **13**: 261-277.

Moriyama, E.N. and Powell, J.R. 1997. Synonymous substitution rates in *Drosophila*: mitochondrial versus nuclear genes. *J Mol Evol* **45**: 378-391.

Mount, D.W. 2004. *Bioinformatics: Sequence and Genome Analysis.* Second Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Mourrain, P., Beclin, C., Elmayan, T., Feuerbach, F., Godon, C., Morel, J.B., Jouette, D., Lacombe, A.M., Nikic, S., Picault, N., *et al.* 2000. Arabidopsis SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance. *Cell* **101**:533-542.

Mozer, B. A., and Benzer, S. 1994. Ingrowth by photoreceptor axons induces transcription of a retrotransposon in the developing Drosophila brain. *Development* **120**: 1049-1058.

Myers, E.W. et al. 2000. A whole-genome assembly of Drosophila. *Science* **287**: 2196-2204.

Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**:443-453.

Nekrutenko, A., and Li, W.H. 2001. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet.* **17**: 619-621.

Nikolajczyk, B.S., Nelsen, B. and Sen, R. 1996. Precise alignment of sites required for mu enhancer activation in B cells. *Mol Cell Biol* **16**: 4544-4554.

Nobrega, M.A. and Pennacchio, L.A. 2003. Comparative genomic analysis as a tool for biological discovery. *J Physiol* **554**: 31-39.

Nuzhdin, S.V. and Mackay, T.F.C. 1995. The genomic rate of transposable element movement in *Drosophila melanogaster*. *Mol. Biol. Evol.* **12**: 180-181.

O'Hagan A, Forster J, 2004. *Bayesian Inference*, volume 2B of *Kendall's Advanced Theory of Staistics*, chapter 9. Arnold, London, 2nd edition.

Ohno, T. 1972. So much "junk" DNA in our genome. Pp. 366-370. In *Evolution of Genetic Systems*. Vol. 23, Brookhaven Symp. Biol.

Ohta, T. 1992. The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics* **23**: 263-286

Ohta, T. 2002. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl. Acad. Sci. U.S.A.* **99**: 16134-37.

Orgel, L.E. and Crick, F.H.C. 1980. Selfish DNA: the ultimate parasite. *Nature* **284**: 604-607.

Pardue, M.L., and DeBaryshe, P.G. 2003. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annu Rev Genet* **37**: 485-511.

Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* **4**: 2444-2448

Pennacchio, L.A. and Rubin, E.M. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* **2**: 100-109.

Petes, T.D., Malone, R.E. and Symington, L.S. 1991. Recombination in yeast. pp. 407-521. In: Broach, J., Jones, E. and Pringle, J. (eds.) *The molecular and cellular biology of the yeast Saccharomyces: genome dynamic, protein synthesis and energetics.* Vol. I, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Petrov, D.A., Lozovskaya, E.R. and Hartl, D.L. 1996. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**: 346-349.

Petrov, D. A., and Hartl, D. L. 1999. Patterns of nucleotide substitution in Drosophila and mammalian genomes. *Proc Natl Acad Sci U S A* **96**: 1475-1479.

Pollard, D.A., Bergman, C.M., Stoye, J., Celniker, S.E. and Eisen, M.B. 2004. Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* **5**: 6.

Presgraves, D. C., 2005. Recombination enhances protein adaptation in Drosophila melanogaster. *Curr Biol* **15**: 1651-1656.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. 1992. *Numerical recipes in C: the art of scientific computing*, chapter 10. Cambrige University Press, Cambridge, UK.

Pyatkov, K. I., Shostak, N. G., Zelentsova, E. S., Lyozin, G. T., Melekhin, M. I., *et al.*, 2002. Penelope retroelements from Drosophila virilis are active after transformation of Drosophila melanogaster. *Proc Natl Acad Sci U S A* **99**: 16150-16155.

Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., *et al.*, 2005. Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol* **1**: 166-175.

Ranganathan, G., Vu, D. and Kern, P.A. 1997. Translational Regulation of Lipoprotein Lipase by Epinephrine Involves a Trans-acting Binding Protein Interacting with the 3' Untranslated Region. *J Biol Chem* 272: 2515-2519.

Rashkova, S., Karam, S.E. and Pardue, M.-L. 2002. Element-specific localization of Drosophila retrotransposon Gag proteins occurs in both nucleus and cytoplasm. *Cell Biology* **99**: 3621-3626.

Reese, J.T. and Pearson, W.R. 2002. Empirical determination of effective gap penalties for sequence comparison. *Bioinformatics* **18**: 1500-1507.

Richards, S. et al. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Res.* **15**: 1-18.

Rizzon, C., Marais, G., Gouy, M. and Biémont, C. 2002. Recombination rate and the distribution of transposable elements in the Drosophila melanogaster genome. *Genome Res.* **12**: 400-407.

Rocha, E.P., Matic, I. and Taddei, F. Over-representation of repeats in stress response genes: a strategy to increase versatility under stressful conditions? *Nucleic Acids Res.* **30**(9): 1886-94.

Sandell, L.L. and Zakian, V.A. 1994. Loss of a yeast telomere: arrest, recovery, and chromosome loss. *Cell* **75**: 729-739.

Sankoff, D., Morel, C. and Cedergren, R.J. 1973. Evolution of 5S RNA and the nonrandomnes of base replacement. *Nature New Biology* **245**: 232-234.

Sankoff, D. and Cedergren, R.J. 1983. Simultaneous comparison of three or more sequences related by a tree. In Sankoff, D. and Kruskal, J.B., eds., *Time Warps, String Edits, and*

Macromolecules: The Theory and Practice of Sequence Comparison. Addison-Wesley. Chapter 9, pp.253-264.

SanMiguel, P.A., Tikhonov, A. and Jin. Y.K. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**: 765-768.

Shabalina, S.A. and Kondrashov, A.S. 1999. Pattern of selective constraint in C. elegans and C. briggsae genomes. *Genet. Res.* **74**: 23-30.

Shapiro, J. A., Huang, W., Zhang, C., Hubisz, M. J., Lu, J., Turissini, D. A., Fang, S., Wang, H. Y., Hudson, R. R., Nielsen, R., Chen, Z., and Wu, C. I. 2007. Adaptive genic evolution in the *Drosophila* genomes, *Proc Natl Acad Sci USA* **104**(7): 2271-6.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**(8): 1034-50.

Singh, N. D., Arndt, P. F., and Petrov, D. A. 2005a. Genomic heterogeneity of background substitutional patterns in Drosophila melanogaster. *Genetics* **169**: 709-722.

Singh, N. D., Davis, J. C. and Petrov, D. A. 2005b. Codon bias and noncoding GC content correlate negatively with recombination rate on the Drosophila X chromosome. *J Mol Evol.* **61**: 315-324.

Singh, N. D., and Petrov, D. A. 2004. Rapid sequence turnover at an intergenic locus in Drosophila. *Mol Biol Evol* **21**: 670-680.

Slawson, E. E., Shaffer, C. D., Malone, C. D., Leung, W., Kellmann, E., *et al.*, 2006. Comparison of dot chromosome sequences from D. melanogaster and D. virilis reveals an enrichment of DNA transposon sequences in heterochromatic domains. *Genome Biol* **7**: R15.

Slotkin, R.K. and Martienssen, R. 2007. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Gen.* **8**: 272-285.

Smale, S.T., and Kadonaga, J,T. 2003. The RNA polymerase II core promoter. *Annu Rev Biochem.* **72**: 449-479.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147**:195-197.

Spradling, A.C., Stern, D.M., Kiss, I., Roote, J., Laverty, T. and Rubin, G.M. 1995. Gene disruptions using P transposable elements: an intergral component of the Drosophila genome project. *Proc. Natl. Acad. Sci. U.S.A.* **92**: 10824-30.

Sparrow, A.H., Price, H.J. and Underbrink, A.G. 1972. A survey of DNA content per cell and per chromosome of prokaryotic and eukaryotic organisms: Some evolutionary considerations. *Brookhaven Symp. Biol.* **23**: 451-494.

Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* doi:10.1038/nature06340.

Takahata, N., Satta, Y. and Klein, J. 1995. Divergence time and population size in the lineage leading to modern humans. *Theor. Popul. Biol.* **48**: 198-221.

Tanimoto K, Yoshida E, Mita S, Nibu Y, Murakami K, Fukamizu A. 1996. Human activin betaA gene. Identification of novel 5' exon, functional promoter, and enhancers. *J Biol Chem* **271**: 32760-32769.

Thomas, C.A. 1971. The genetic organization of chromosomes. *Annual Review of Genetics* **5**: 237-256.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., et al. 2003. Comparative analysis of multi-species sequences from targeted genomic regions. *Nature* **424**: 788-793.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994. CLUSTAL W-Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673-4680.

Thornburg, B.G., Gotea, V., and Makalowski, W. 2006. Transposable elements as a significant source of transcription regulating signals. *Gene* **365**: 104-110.

Thorne, J.L., Kishino, H. and Felsenstein, J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33**: 114-124.

Thorne, J.L., Kishino, H. and Felsenstein, J. 1992. Inching toward reality-An improved likelihood model of sequence evolution. *J. Mol. Evol.* **34**: 3-16.

Tiffany HL, Handen JS, Rosenberg HF. 1996. Enhanced expression of the eosinophil-derived neurotoxin ribonuclease (RNS2) gene requires interaction between the promoter and intron. *J Biol Chem* **271**: 12387-12393.

Ting SJ. 1995. A binary model of repetitive DNA sequence in *Caenorhabditis elegans*. *DNA Cell Biol.* **14**: 83-85.

Trifonov, E.N. 1989. The multiple codes of nucleotide sequences. *Bull. Math Biol.* **51**: 417-432.

van de Lagemaat, L. N., Landry, J. R., Mager, D. L. and Medstrand, P. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* **19**: 530-536.

Vandendries ER, Johnson D, Reinke R. 1996. Orthodenticle is required for photoreceptor cell development in the *Drosophila* eye. *Dev Biol* **173**: 243-255.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., *et al.*, 2001. The sequence of the human genome. *Science* **291**: 1304-1351.

Vieira, C. and Biémont, C. 2004. Transposable element dynamics in two sibling species: *Drosophila melanogaster* and *Drosophila simulans*. *Genetica* **120**:115-123.

Voytas, D.F. and Boeke, J.D. 2002. *Mobile DNA II*, Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M. Eds. American Society for Microbiology, Washington, DC.

Wang, J., Keightley, P. D. and Johnson, T. 2006. MCALIGN2: faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. *BMC Bioinformatics* **7**: 292.

Wang, J., Keightley, P.D., and Halligan, D.L. 2007. Effect of divergence time and recombination rate on molecular evolution of *Drosophila* INE-1 transposable elements and other candidates for neutrally evolving sites. *J Mol Evol* **65**: 627-639.

Waterston, R. and Sulston, J. 1995. The genome of Caenorhabditis elegans. *Proc. Natl. Acad. Sci. USA* **92**: 10836-40.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., *et al.*, 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics* **16**: 97-159.

White, S. E., Habera, L. F. and Wessler, S. R. 1994. Retrotransposons in the flanking regions of normal plant genes: a role for copia-like elements in the evolution of gene structure and expression. *Proc Natl Acad Sci U S A* **91**: 11792-11796.

Yang, H.-P., Hung, T.-L., You, T.-L. and Yang, T.-H. 2006 Genomewide comparative analysis of the highly abundant transposable element *DINE-1* suggests a recent transpositional burst in *Drosophila yakuba*. *Genetics* **173**(1): 189–196.

Yang, H.-P. and Barbash, D.A. 2008. Abundant and species-specific DINE-1 transposable elements in 12 Drosophila genomes. *Genome Biology* **9**: R39.

Yang, Z. 1997. On the estimation of ancestral population sizes. *Genet. Res.* **69**: 111-116.

Yang, Z., Boffelli, D., Boonmark, N., Schwartz, K. and Lawn, R. 1998. Apolipoprotein(a) gene enhancer resides within a LINE element. *Journal of Biological Chemistry* **273**: 891-897.

Yi, S., Summers, T. J., Pearson, N. M. and Li, W.-H. 2004. Recombination Has Little Effect on the Rate of Sequence Divergence in Pseudoautosomal Boundary 1 Among Humans and Great Apes. *Genome Res.* **14**: 37 – 43.

Zhang, Z.L. and Gerstein, M. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Research* 31: 5338-5348.

Zhu, J., Liu, J.S. and Lawrence, C.E. 1998. Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14**: 25-39.