STUDIES ON THE SIGNIFICANCE OF LTR-RETROTRANSPOSONS ON GENE STRUCTURE AND FUNCTION

by

ANDREA M. McCOLLUM

(Under the direction of John F. McDonald)

ABSTRACT

Transposable elements are mobile DNA sequences that can move from one chromosome position to another. These elements are ubiquitous in eukaryotes and compose a large percentage of some eukaryotic genomes. The current thesis considers the impacts of transposable elements on host genes.

Four LTR-retrotransposon/gene associations in *Caenorhabditis elegans* have been surveyed across known populations of the species. All four associations are detected in high frequency or fixed throughout the species.

In another population study, a euchromatic LTR-retrotransposon/gene association in *Drosophila melanogaster* is only detected in one population while a heterochromatic association is detected in all populations examined. Sequence analysis of the LTR sequence suggests that selection is maintaining this retrotransposon sequence in the *Drosophila melanogaster* species as well as a sister taxa, *D. mauritiana*.

INDEX WORDS: Genetics, Transposable element, Heterochromatin, Molecular evolution, *Drosophila*, *Caenorhabditis*

STUDIES ON THE SIGNIFICANCE OF LTR-RETROTRANSPOSONS ON GENE STRUCTURE AND FUNCTION

by

ANDREA M. McCOLLUM

B.S., The University of Georgia, 1999

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2002

STUDIES ON THE SIGNIFICANCE OF LTR-RETROTRANSPOSONS ON GENE STRUCTURE AND FUNCTION

by

ANDREA M. McCOLLUM

Approved:

Major Professor: John F. McDonald

Committee: Wyatt W. Anderson

Rodney Mauricio

Electronic Version Approved:

Gordham L. Patel

Dean of the Graduate School

The University of Georgia

August 2002

ACKNOWLEDGEMENTS

I would like to thank my advisor, John F. McDonald, for his support and encouragement here at UGA and for future endeavors. I thank my committee members, Wyatt Anderson and Rodney Mauricio, for their encouragement and words of advice. I am indebted to past and present members of the McDonald lab for technical assistance, support, and laughter. In particular, thanks to Nathan Bowen, Kevin Fielman, Eric Ganko, Lilya Matyunina, and Nina Schubert. Special thanks to members of the Anderson lab (especially Yong Kyu Kim) for teaching me all I know about flies and for that extra push I needed to go to graduate school.

Thanks to my family for support and for actually showing interest in worms and flies. I could not have gotten through this degree without the patience, understanding, and encouragement of Eric Caldwell. Even though you have no interest in worms and flies you have tried to understand a small fraction of biology. Basically, thanks to family and friends for caring.

There are many more people that are not named here that I have learned from and enjoyed being a small part of their lives. Thanks to you all!

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Transposable elements (TEs) are mobile DNA sequences able to move from one chromosome position to another. TEs are classified based on the mechanism by which they transpose, and there are two major classes of TEs. Class I elements, retroelements, transpose by the reverse transcription of an RNA intermediate [1]. Class II elements, DNA elements, transpose from DNA to DNA directly by the enzyme transposase [1].

The existence of TEs were first inferred by Barbara McClintock in 1948 in maize [2]. McClintock showed that mobile elements could regulate or change host gene expression and because of this ability, McClintock called these elements "controlling elements" [2, 3]. She also showed that mobile elements are activated when the genome is under shock or stress, and these results led McClintock to propose that these elements could have a role in genome restructuring [3]. This restructuring might give advantages to the host genome; therefore, the genome would favor the presence of these elements [4].

The impacts TEs could have for genomes and genome evolution were quite obvious to McClintock, however the evolutionary significance of TEs was stifled by the 'Selfish DNA' theory of TEs [3, 5, 6]. In the early 1980s two papers argued that the spread of TEs in a genome is a result of the self-replication of these selfish elements [5, 6]. Another author backed up the theory by theoretically showing that TEs could be maintained in populations even while imparting a slight disadvantage to their host [7]. Although there was little support within the scientific community to study the

evolutionary consequences of TEs, molecular data began to show ways in which TEs may transform or impact genomes.

Early research on TEs revealed that these mobile sequences are components of both prokaryotic and eukaryotic genomes [8, 9]. Greater than 50% of the maize genome and an estimated 43% of the human genome is composed of TEs [10, 11]. Given the high percentage of TEs in genomes, one might expect that a significant portion of genes would contain TEs. A genomic analysis of the sequenced human genome revealed that 4% of human genes have retrotransposon sequences contained within protein-coding regions [12]. It has been projected that 4-6% of genes in *Arabidopsis thaliana* will contain retrotransposons [13].

TE insertions into the coding regions of genes were generally thought to be deleterious and quickly eliminated from the population; however, TEs are parts of the coding regions of a large number of genes [12, 14]. Full-length and fragmented TEs can be functional components of host genes (Fig 1.1) [15]. LTR-retrotransposons contain two LTRs, which contain critical promoter, enhancer, and polyadenylation sequences for the element, and there are documented instances where these sequences have contributed to host gene function and regulation (e.g., [16-18]).

In addition to being functional components of genes, TEs are also structural components of heterochromatin in *Drosophila* [19]. Heterochromatin consists of highly repeated and middle repetitive sequences [20, 21]. The middle repetitive sequences are seen as descendents of active TEs that inserted into heterochromatin, which is thought of as being transcriptionally inactive (e.g., [22, 23]). Although heterochromatin was thought of as a junkyard of the genome, functionally important genes have been mapped to this

area of the genome (e.g., [21, 24-30]). *Drosophila* heterochromatic genes are associated with retrotransposon sequences, and it has been suggested that TEs may locally alter chromatin structure [30-35]. Therefore, perhaps the TEs located within heterochromatin are contributing to the ability of heterochromatic genes to be expressed.

The following two chapters will focus on six TE/gene associations in *Caenorhabditis elegans* and *Drosophila melanogaster*. The experiments conducted are some in a series of experiments designed to evaluate the possible adaptive significance of TE insertions in host genes

References

1. Finnegan DJ: **Transposable elements**. *Curr Opin Genet Dev* 1992, **2**:861-7.
2. McClintock B: **Mutable loci in maize**. *Carnegie Institute of Washington Year Book* 1948, **47**:155-169.
3. McClintock B: **The Significance of Responses of the Genome to Challenge**. *Science* 1984, **226**:792-801.
4. Wessler SR: **Turned on by stress. Plant retrotransposons**. *Curr Biol* 1996, **6**:959-61.
5. Doolittle WF, Sapienza C: **Selfish genes, the phenotype paradigm and genome evolution**. *Nature* 1980, **284**:601-3.
6. Orgel LE, Crick FH: **Selfish DNA: the ultimate parasite**. *Nature* 1980, **284**:604-7.
7. Hickey DA: **Selfish DNA: a sexually-transmitted nuclear parasite**. *Genetics* 1982, **101**:519-31.

8. Jordan MA, Saedler H, Starlinger P: **Strong-polar mutations in the transferase gene of the galactose operon**. *E. coli Molecular and General Genetics* 1967, **100**:296-306.

9. Finnegan DJ: **Repeated gene families in *Drosophila melanogaster***. *Cold Spring Harbor Symposia on Quantitative Biology* 1977, **42**:1053-1063.

10. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL: **Nested retrotransposons in the intergenic regions of the maize genome**. *Science* 1996, **274**:765-8.

11. Li WH, Gu Z, Wang H, Nekrutenko A: **Evolutionary analyses of the human genome**. *Nature* 2001, **409**:847-9.

12. Nekrutenko A, Li WH: **Transposable elements are found in a large number of human protein- coding genes**. *Trends Genet* 2001, **17**:619-21.

13. Kumar A, Bennetzen JL: **Plant retrotransposons**. *Annu Rev Genet* 1999, **33**:479-532.

14. Brosius J: **Genomes were forged by massive bombardments with retroelements and retrosequences**. In: *Transposable Elements and Genome Evolution*. pp. 209-238; 2000: 209-238.

15. McDonald JF: **Transposable elements: possible catalysts for organismic evolution**. *Trends in Ecology and Evolution* 1995, **10**.

16. Banville D, Boie Y: **Retroviral long terminal repeat is the promoter of the gene encoding the tumor-associated calcium-binding protein oncomodulin in the rat**. *J Mol Biol* 1989, **207**:481-90.

17. Stavenhagen JB, Robins DM: **An ancient provirus has imposed androgen regulation on the adjacent mouse sex-limited protein gene**. *Cell* 1988, **55**:247-54.

18. Mager DL, Hunter DG, Schertzer M, Freeman JD: **Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (HHLA2 and HHLA3)**. *Genomics* 1999, **59**:255-63.

19. Pimpinelli S, Berloco M, Fanti L, Dimitri P, Bonaccorsi S, Marchetti E, Caizzi R, Caggese C, Gatti M: **Transposable elements are stable structural components of Drosophila melanogaster heterochromatin**. *Proc Natl Acad Sci U S A* 1995, **92**:3804-8.

20. Appels R, Peacock WJ: **The arrangement and evolution of highly repeated (satellite) DNA sequences with special reference to Drosophila**. *Int Rev Cytol Suppl* 1978, **Suppl**:69-126.

21. Brutlag DL: **Molecular arrangement and evolution of heterochromatic DNA**. *Annu Rev Genet* 1980, **14**:121-44.

22. Charlesworth B, Langley CH: **The population genetics of Drosophila transposable elements**. *Annu Rev Genet* 1989, **23**:251-87.

23. Laurie CC, Bridgham JT, Choudhary M: **Associations between DNA sequence variation and variation in expression of the Adh gene in natural populations of Drosophila melanogaster**. *Genetics* 1991, **129**:489-99.

24. Hilliker AJ, Holm DG: **Genetic analysis of the proximal region of chromosome 2 of Drosophila melanogaster. I. Detachment products of compound autosomes**. *Genetics* 1975, **81**:705-21.

25. Hilliker AJ: **Genetic analysis of the centromeric heterochromatin of chromosome 2 of Drosophila melanogaster: deficiency mapping of EMS-induced lethal complementation groups**. *Genetics* 1976, **83**:765-82.

26. Ganetzky B: **On the components of segregation distortion in Drosophila melanogaster**. *Genetics* 1977, **86**:321-55.

27. Pimpinelli S, Sullivan W, Prout M, Sandler L: **On biological functions mapping to the heterochromatin of Drosophila melanogaster**. *Genetics* 1985, **109**:701-24.

28. Marchant GEaDGH: **Genetic analysis of the heterochromatin of chromosome 3 in *Drosophila melanogaster*. I. Products of compound autosome detachment.** *Genetics* 1988a, **120**:503-517.

29. Marchant GEaDGH: **Genetic analysis of the heterochromatin of chromosome 3 in *Drosophila melanogaster*. II. Vital loci identified through EMS mutagenesis.** *Genetics* 1988b, **120**:519-532.

30. Devlin RH, Holm DG, Morin KR, Honda BM: **Identifying a single-copy DNA sequence associated with the expression of a heterochromatic gene, the light locus of Drosophila melanogaster**. *Genome* 1990, **33**:405-15.

31. Shapiro J: **DNA insertion elements and the evolution of chromosome primary structure**. *Trends Biochem. Sci.* 1977, **2**:622-627.

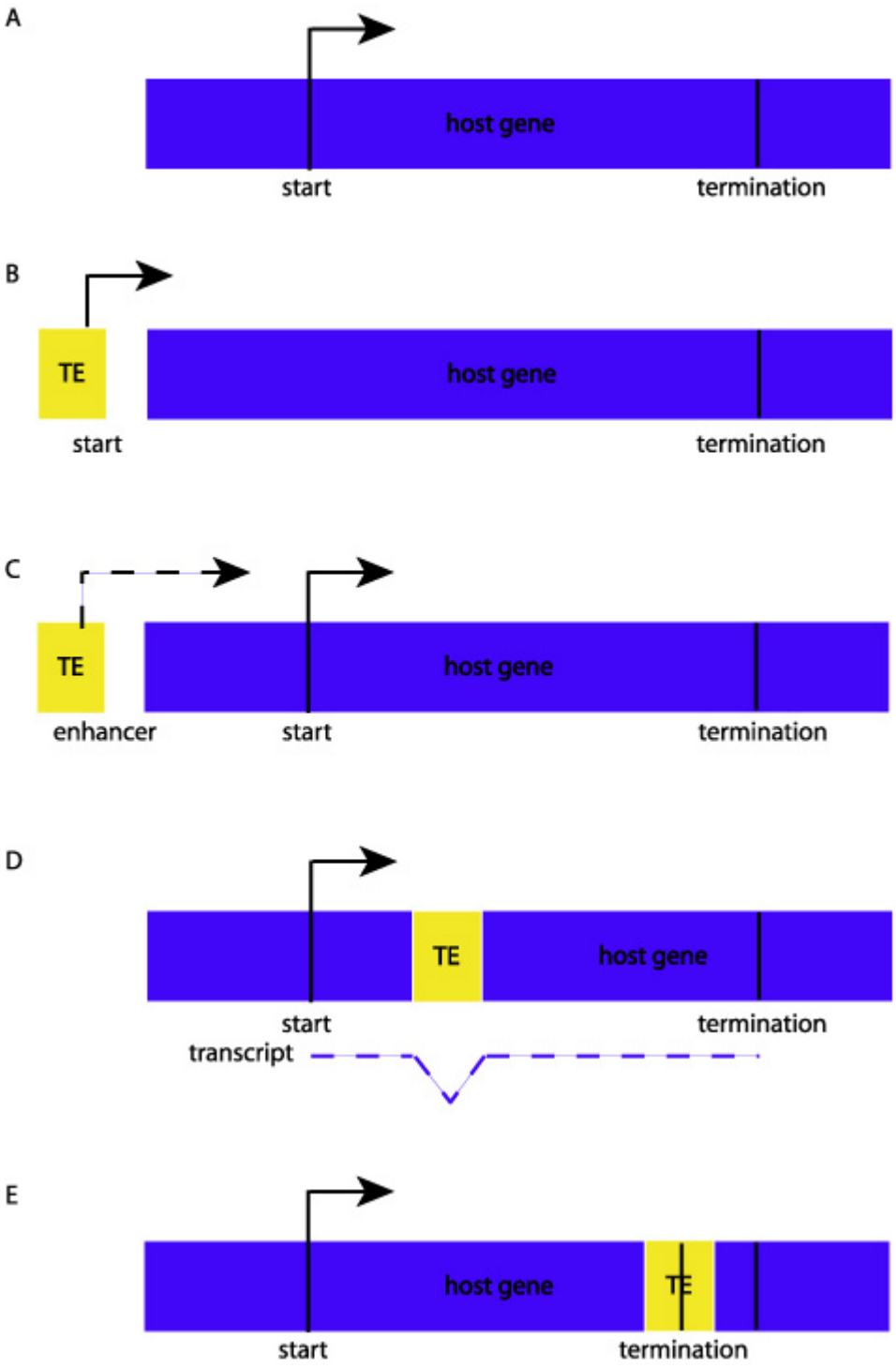
32. Dimitri P, Junakovic N: **Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin**. *Trends Genet* 1999, **15**:123-4.

33. Berghella L, Dimitri P: **The heterochromatic rolled gene of Drosophila melanogaster is extensively polytenized and transcriptionally active in the salivary gland chromocenter**. *Genetics* 1996, **144**:117-25.

34. Dimitri P, Junokovic, N., and B. Arca: **Nested transposons within the intron II of rolled, a heterochromatic gene of *Drosophila melanogaster***. *Dros Res Conf* 1999, **40**:663B.

35. Gatti M, Pimpinelli S: **Functional elements in Drosophila melanogaster heterochromatin**. *Annu Rev Genet* 1992, **26**:239-75.

Figure 1.1: **Possible effects of a TE insertion on gene expression.** (A) A typical host gene without a TE insertion. The arrow indicates the transcriptional start site and the solid line indicates the transcriptional stop site (termination). (B) Insertion of a TE 5' to a gene resulting in a new transcriptional start site provided by the TE. (C) Insertion of a TE 5' to a gene resulting in a new enhancer provided by the TE. A new enhancer can also occur if a TE is contained within a gene or 3' to a gene. (D) Insertion of a TE within a gene. The TE provides new splice signals resulting in an altered splicing pattern (in comparison to the wild type gene, (A)). (E) Insertion of TE within a gene resulting in a new termination sequence provided by the TE.


A
host gene
start
termination
B
TE
host gene
start
termination
C
TE
host gene
enhancer
start
termination
D
TE
host gene
start
termination
transcript
E
host gene
TE
start
termination

CHAPTER 2

DISTRIBUTION OF TE / GENE ASSOCIATIONS IN NATURAL POPULATIONS OF *C. ELEGANS*

Background

Since the discovery of TEs, they have been found in significant numbers within the genomes of most living organisms. For example, greater than 50% of the maize genome is comprised of TEs and an estimated 43% of the human genome is composed of TEs [1], [2]. For many decades, TEs were generally thought of as purely selfish DNA (e.g.,[3, 4]). More recently scientists have shifted their attention to the impact TEs have on the host genome and the possible coadaptation of elements and their hosts (e.g., [5-8]).

A TE can be maintained in populations even while imparting a slight disadvantage to their host [3, 4, 9]; but, a TE may also be maintained in host genomes because of adaptive significance the TE confers upon the host (e.g., [10]). Insertions of TEs into coding regions of genes were generally thought to be deleterious and eliminated from the population quickly. However, TEs have been found to be part of the coding regions of a large number of genes [11, 12]. A recent genomic analysis of the human genome (13,799 genes) showed that 4% of human genes have retrotransposon sequences contained within protein-coding regions [12], and it has been projected that 4-6% of genes in *Arabidopsis thaliana* will contain retrotransposons [13].

Full-length and fragmented TEs have been shown to be functional and critical parts of host genes and to contribute to host gene regulatory variation (e.g., [14, 15]). Inserted LTRs of LTR-retrotransposons and retroviruses contain critical promoter,

enhancer, and polyadenylation sequences for the retroelement and have been shown to contribute to host gene function and regulation (e.g., [16-18]). There is evidence that a retrotransposon insertion contained within an intron of *Cht3*, a heterochromatic gene in *Drosophila*, is of adaptive significance to the host genome [10].

The availability of genome sequence data has given researchers an unprecedented opportunity to study the impacts of TEs on host genomes (e.g., [12, 19, 20]). The complete genome sequence of the nematode *Caenorhabditis elegans* has allowed researchers to identify the majority of full-length and fragmented LTR-retrotransposons in the genome, *Cer* elements (*C. elegans* retroelements) [20, 21]. Ganko, Fielman, & McDonald identified four LTR-retrotransposon/gene associations in which the LTR-retrotransposons map to putative splice sequences and termination regions suggesting that these *Cer* elements may contribute to gene expression and evolution [20]. Here we continue a series of experiments designed to establish if the identified TE/gene associations are widespread in *C. elegans* natural populations.

Results

1. **The genes and *Cer* elements examined in this study are detectable by PCR in all *C. elegans* populations surveyed.**

   PCR analysis was conducted on four genes (6R55.2, C56G3.2, F20B4.6, and F20B4.6) and four LTR-retrotransposons (*Cer16-2, Cer9, Cer2,* and *Cer16-2*) to determine if the genes and LTR-retrotransposons are present in 14 *C. elegans* populations and can be detected by our PCR analysis. A large number of genes (~50%) in the *C. elegans* genome are uncharacterized, and these genes are

represented by clone-specific gene numbers [22]. PCR primers were designed to amplify coding regions of the four genes and to amplify element-specific sequences of the four elements (Figure 1.1). The PCR primers were designed from the *C. elegans* genome sequence, strain N2. Performing the PCR analysis to detect the genes and elements across different populations verified that the primers worked in DNA samples representing all populations examined in this study. Expected PCR products were produced in all *C. elegans* populations for the four genes and the four elements.

2. The presence of the *Cer16-2* insertion into 6R55.2, the *Cer9* insertion into C56G3.2, and the *Cer2* insertion into F20B4.6 are variable across *C. elegans* populations surveyed.

PCR analysis was conducted on the four element/gene associations to determine if the insertions have been maintained across 14 *C. elegans* populations. The 3' LTR of a full-length *Cer16-2* element is contained within the 5' end of 6R55.2, a putative gene with unknown function (Figure 2.1). A *Cer9* LTR from a fragmented element is positioned in the 5' end of C56G3.2, an aldo/keto reductase homolog (Figure 2.1). A *Cer2* solo LTR is contained within the 3' region of F53E10.5, a gene with homology to a helicase domain (Figure 2.1). PCR primers were designed from the *C. elegans* genome sequence (strain N2) to amplify regions of the genes and respective elements. For example, a forward primer for *Cer9* and a reverse primer for C56G3.2 were used to amplify the insertion of *Cer9* contained in C56G3.2. The *Cer9* insertion was not detected in the Vancouver and Hawaii populations (Table 2.1). The full-length

Cer16-2 element insertion was not detected in the Adelaide and Vancouver populations, and the *Cer2* solo LTR insertion was not detected in the Madison, Vancouver, and Hawaii populations (Table 2.1).

3. The *Cer16-1* insertion contained within F20B4.6 was detected in all populations surveyed.

A *Cer16-1* solo LTR is positioned entirely within an intron of gene F20B4.6, a member of the glucosyltransferase family (Figure 2.1). As is described in the previous results for the PCR analysis of the *Cer9, Cer16-2,* and *Cer2* insertions, appropriate gene and element primers were used to amplify the *Cer16-1* insertion within the F20B4.6 gene. This insert contained within an intron was in detected in all 14 populations surveyed (Table 2.1, Figure 2.2).

4. The genes, elements, and associations examined in this study were not detected in *C. briggsae* or *C. remanei.*

PCR analysis on the four genes, elements, and insertions as described above was conducted on populations of *Caenorhabditis briggsae* (strain AF16) and *Caenorhabditis remanei* (strain PB206), sister taxa of *Caenorhabditis elegans.* The sister taxa are expected to have diverged 25-50 million years ago [23]. No detectable product was produced for the four genes, elements, or element/gene associations for these two taxa.

The genome of *C. briggsae* is currently being sequenced, and new sequence information is available and is continuously updated on www.wormbase.org. This

sequence data is searchable, and a BLASTN search was conducted on the *C. briggsae* genome data with sequences from all four genes and elements studied. As of 6/12/02, there was no *C. briggsae* sequence with significant similarity to F53E10.5 or 6R55.2. There was similarity to F20B4.6, with low similarity in regions were primers were used. Also, there was similarity to C56G3.2.

5. **There was no detectable expression for any of the four genes examined using the RNA blot technique.**

   Varying concentrations of total genomic DNA and total RNA from strains in which certain TE/gene associations were not detected (AB2, TR403, KR314, and CB4856) and a control strain (N2) was bound to a nylon membrane via a slot blot manifold (Schleicher & Schuell, Inc., Keene, NH). This membrane was then hybridized to each gene (6R55.2, C56G3.2, F20B4.6, and F20B4.6) to detect for differences in expression between the control strain that contains the insert of interest (N2) and strains in which inserts were not detected. For each gene, there was no detectable hybridization to RNA. However, there was detectable hybridization to DNA with increasing intensity of hybridization corresponding to increasing amounts of DNA. When the membrane was hybridized to an 18s ribosomal DNA control there was hybridization to both RNA and DNA with increasing intensity of hybridization corresponding to increasing amounts of DNA and RNA.

Discussion

The purpose of the study presented here was to determine if TE/gene associations identified in the *C. elegans* N2 strain are widespread in natural populations. The establishment of the possible adaptive significance is a three-stage process. Stage I is the identification of an element/gene association. An association is defined as an element sequence located within or adjacent to a gene, and these associations are identified by computational analysis of genome databases. Stage II consists of a series of analyses designed to determine the extent to which the identified associations is polymorphic in natural populations. Stage III consists of molecular and possibly more detailed computational analyses designed to determine the functional consequences of a TE insertion on a host gene.

As stated above, Stage I is the identification of TE/gene associations in a genome. The identified element should have the potential to contribute to gene structure and/or function; therefore, the identified TE should be within the gene or near enough to exert a regulatory effect. In *C. elegans* most gene regulatory sequences are located within 1000bp of the gene coding sequence [24]. Stage II of the process is a comparative analysis of the TE/gene associations within populations of a species and between species. The identification of a TE/gene association in a sequenced genome is not necessarily representative of a species nor does this one identification yield insights into the evolutionary history of a particular TE/gene association. The identification may represent a strain-specific mutation that is not typical of the species. Stage II analysis of the distribution of TE/gene associations within and among species may provide valuable insight into the evolutionary history of the association and its possible adaptive

significance. Stage III of the analysis uses molecular and/or other appropriate techniques to determine the functional significance and consequences of a TE/gene association. A non-adaptive TE/gene association may drift to high frequency by random processes over time [25]; therefore, the comparative analyses conducted in Stage II cannot yield conclusive results about the adaptive significance of a TE. Functional and/or other appropriate tests can yield additional support for the adaptive hypothesis and help prove or disprove neutrality of the insertion.

The current study is a continuation of a series of experiments to determine the distribution of TE/gene associations in *C. elegans* and their possible adaptive significance. Stage I, identification of TE/gene associations in a genome, was completed by Ganko, Fielman, & McDonald [20]. *C. elegans* is a nematode whose genome has been completely sequenced, and access to genome sequences has greatly facilitated identification of TE/gene associations. Ganko, Fielman, & McDonald initially examined the four gene/element associations assayed here by blasting the *C. elegans* EST database for homology to *Cer* LTR sequences [20]. These ESTs were then mapped to putative genes or genes that belong to supported gene families [20]. In addition to the basic requirements for identifying a TE/gene association in Stage I, Ganko, Fielman, & McDonald showed that the four *Cer* element insertions map to putative splice sequences and termination regions of the genes [20]. Ganko, Fielman, & McDonald also showed by RT-PCR that the four te/gene associations are transcribed [20].

A Stage II, comparative analysis, was conducted in the present study. We have shown that three element/gene associations in *C. elegans* are variable across 14 populations. One association, *Cer16-1* within the intron of F20B4.6, was detected in all

populations tested. The four genes, elements, or associations were not detected in closely related *C. briggsae* or *C. remanei*. The following discussion focuses on the Stage II analysis.

C. elegans is a species of nematodes that is widely dispersed throughout the world [26, 27]. Previous studies attempting to type different populations by Tc1, a DNA transposon, or other molecular markers have shown that different races do exist in the wild [28, 29]. Some races are endemic existing in only one location (example: Hawaiian race) while others are more widespread (example: a Bristol race has been recovered in Wisconsin and California) [29]. Since different races can be isolated from the same location, there is probably very little gene flow between *C. elegans* races, and populations are probably highly inbred [27]. In this study we surveyed 14 populations from 13 known distinctive races, and the populations used are representative of the isolates available to researchers [27].

If a TE insertion is deleterious or contributes to a significant loss in fitness the insertion is expected to be lost from the population. Conversely, if the insertion leads to increased fitness for the organism, this insertion is expected to be maintained and would eventually be expected to increase frequency or fixed in the species by selection. Certainly, the possibility exists that an individual insertion mutant to drift to high frequency or even become fixed in an individual population, but this is unlikely to occur simultaneously in many separate populations.

Here, three element/gene associations are found in high frequency in the species *C. elegans* and one association is fixed in *C. elegans*. These results suggest that these four TE/gene associations probably arose prior to the divergence of these races and/or

that the association may be of adaptive significance. Ganko, Fielman, and McDonald showed that the four *Cer* element insertions in this study map to putative splice sequences and termination regions, also suggesting that these particular insertions may contribute to gene regulation [20]. In addition, RT-PCR has shown that transcripts from the TE/gene associations are produced for all four associations in the sequenced strain N2 [20]. All of these results are highly suggestive that these four gene/element associations in *C. elegans* could be of adaptive significance to the host genome. However, a final resolution of this issue will require further molecular analysis.

C. elegans and *C. briggsae* are thought to have diverged from each other approximately 25-50 mya [23]. There are a high number of silent substitutions in *C. briggsae* in comparison to *C. elegans*, and this indicates that *C. briggsae* is quite divergent from *C. elegans* [30]. Sequences of the mitochondrial gene for cytochrome oxidase subunit two and the nuclear gene for calmodulin in *C. elegans, C. briggsae,* and *C. remanei* indicate a possible trichotomous branching of the *Caenorhabditis* clade for these three species [31]. The four genes, elements, and associations assayed in this study were not detected in *C. briggsae* or *C. remanei*. We cannot make a conclusion about the evolutionary history of these four element/gene associations in the *Caenorhabditis* clade because there were no detectable PCR products.

Stage III of the analysis, determining the functional consequence of a TE insertion, was begun in this study. Unfortunately, RNA blot analysis did not detect expression for any gene in all the strains tested (including N2). These results are inconsistent from RT-PCR results from Ganko, Fielman, & McDonald showing that the four loci in question are transcribed in the N2 strain. The RNA blot analysis is much less

sensitive than quantitative RT-PCR to detect low abundance or single copy transcripts. Future experiments using quantitative RT-PCR should be done to adequately assess the functional significance of the TE insertions into host genes.

Materials and Methods

C. elegans stock maintenance

All *C. elegans* stocks except for NS3598 were obtained from the *Caenorhabditis elegans* Genetics Center, University of Minnesota. NS3598 was kindly received from Michael Ailion, University of Washington-Seattle. All stocks were maintained on mixed staged plates as described in Wood [32].

DNA isolation

Mixed stages of *C. elegans* were washed off a plate with M9 buffer [33], and washed in M9 buffer several times. The worms were washed a final time in 4ml TEN (20mM Tris pH7.5, 50mM EDTA, 100mM NaCl). The worms were digested in 25µl 10% SDS, 2.5µl Proteinase K (20 mg/ml), and 1.0 µl beta-mercaptoethanol at 50-60°C for at least 1 hour or until there was no tissue visible. 3µl RNase A (10mg/ml) was added and the sample was incubated at 37°C for 30 minutes to 1 hour. The resulting solution was extracted once with phenol : chloroform and once with chloroform. After extraction, one volume of isopropanol was added, and the sample was then centrifuged for 10 minutes to pellet nucleic acids. The pellet was washed with 700µl 70% ethanol and centrifuged for 5 minutes on high speed, and then allowed to air dry. The DNA was resuspended in 50µl TE and stored at -20°C.

PCR analysis

PCR primers were designed using MacVector 7.0 (http://www.gcg.com) and synthesized by Integrated DNA Technologies (Coraville, IA). PCR primer sequences are listed in Table 2.2 and shown on Figure 2.1. PCR analysis was conducted for the four genes, elements, and element-gene associations on all strains. The primers used for the analysis are listed in Table 2. The annealing temperatures for primer pairs used in this study are as follows: C56G3.2 (f) – C56G3.2 (r) is 52ºC, *Cer9* (f) – *Cer9*(r) is 53ºC, *Cer9* (f) – C56G3.2 (r) is 55ºC; F53E10.5 (f) - F53E10.5 (r) is 57ºC, *Cer2* (f) – *Cer2* (r) is 51ºC, *Cer2* (f) - F53E10.5 (r) is 54ºC; F20B4.6 (f1) – F20B4.6 (r) is 54ºC, *Cer16-1* (f) – *Cer16-1* (r) is 55ºC, F20B4.6 (f2) - *Cer16-1* (r) is 53ºC; 6R55.2 (f) – 6R55.2 (r) is 54ºC, *Cer16-2* (f) – *Cer16-2* (r) is 54ºC, and *Cer16-2* (f) – 6R55.2 (r) is 58ºC.

The PCR reaction for every set of primers except for *Cer16-2* (f) – 6R55.2 (r) is as follows: 3.0mM $MgCl_2$, 10X PCR buffer supplied by Pierce Endogen (500mM KCl, 100mM Tris-HCl pH 9.0, Rockford, IL), 5% DMSO, 0.2mM dNTPs, 0.5mM each primer, and 0.5U Taq DNA polymerase (Pierce Endogen, Rockford, IL) in a 25µl reaction. The PCR reaction for primer pair *Cer16-2* (f) – 6R55.2 (r) is as follows: 3.0mM $MgCl_2$, Opti-Prime™ 10X buffer #3 (10mM Tris-HCl pH8.3, 3.5mM $MgCl_2$, 25mM KCl, Stratagene, La Jolla, CA), Master Mix™ 50X buffer (20mM Tris-HCl, 250nM EDTA, Stratagene, La Jolla, CA), 5% DMSO, 0.2mM dNTPs, 0.5mM each primer, and 0.5U Taq DNA polymerase (Pierce Endogen, Rockford, IL) in a 25µl reaction.

All PCR reactions were carried out in a Hot Top equipped Robocycler® Gradient Temperature Cycler (Stratagene, La Jolla, CA) using the following program: 1 cycle of 94ºC for 3min; 30 cycles of 94ºC for 30sec, appropriate annealing temperature for 30sec,

68ºC for 1.5min; and 68ºC for 5 min. PCR products were resolved on a 1% agarose gel in 0.5X TBE running buffer containing 0.25µg mL^{-1} ethidium bromide. Gel images were visualized by UV transillumination.

RNA isolation

Mixed staged populations were washed off a plate in dH2O and centrifuged. The pellet was washed twice in dH2O, and frozen in liquid nitrogen. Total RNA was isolated using the guanidinium isothiocyanate: phenol method [34].

Probe production

PCR products produced previously for the detection of the genes C56G3.2, F53E10.5, and 6R55.2 were used as probes. To produce a gene product for F20B4.6 a second set of gene primers were utilized because the first set of gene primers produced a product that contained *Cer16-1*; the second set of primers did not produce a product that contained *Cer16-1*. The primer pair used is F20B4.6 (f3) – *Cer16-1* (r) and the annealing temperature is 53ºC. The PCR reaction for the F20B4.6 gene reaction is 1.5mM $MgCl_2$, 10X PCR buffer supplied by Pierce Endogen (500mM KCl, 100mM Tris-HCl pH 9.0, Rockford, IL), 5% DMSO, 0.2mM dNTPs, 0.5mM each primer, and 0.5U Taq DNA polymerase (Pierce Endogen, Rockford, IL) in a 25µl reaction. 18S rRNA was used as a control. The primers were provided for by Ambion (QuantumRNA™ Universal 18S, Austin, TX). The 18S reaction is the same reaction as described for the genes C56G3.2, F53E10.5, and 6R55.2.

All PCR reactions were carried out in a Hot Top equipped Robocycler® Gradient Temperature Cycler (Stratagene, La Jolla, CA) using the following program: 1 cycle of

94ºC for 3min; 30 cycles of 94ºC for 30sec, appropriate annealing temperature for 30sec, 68ºC for 1.5min; and 68ºC for 5 min. PCR products were resolved on a 1% agarose gel in 0.5X TBE running buffer containing 0.25µg mL^{-1} ethidium bromide.

25ng of each PCR product was nick-translated by using Prime-It® Random Primer Labeling Kit (Stratagene, La Jolla, CA). αP^{32}-dCTP was provided by PerkinElmer Life Sciences, Inc. (Boston, MA). Excess nucleotides were removed with mini Quick Spin DNA columns (Roche, Indianapolis, IN).

RNA Blot

RNA was blotted onto a Hybond N+ membrane (Amersham Pharmacia Biotech, Sunnyvale, CA) by using a Minifold II (Schleicher & Schuell, Inc., Keene, NH). The procedure used for the blot is detailed in Sambrook, Fritsch, and Maniatis [35]. 0µg – approximately 9µg of RNA and 0µg – approximately 7µg of DNA from strains AB2, TR403, CB4856, KR314, and N2 were blotted on the membrane. After transferring DNA and RNA to the membrane, the membrane was UV crosslinked using an Electronic Ultraviolet Crosslinker (Ultra-Lum, Inc., Claremont, CA).

Hybridization

The membrane was prehybridized in 6X SSC, 2X Denhardt's reagent [33], and 0.1% SDS for 1-2 hours at 65°C. The probe was boiled for 10 minutes and added to 30ml of the prehybridization buffer; this was then added to the membrane. The membrane was hybridized at 65°C overnight. After hybridization the membrane was washed three times with 2X SSC, 0.1%SDS for 10-20 minutes at 65°C. These washes were followed by one wash with 0.2X SSC, 0.1% SDS for 10-20 minutes at 65°C and a rinse in 0.2X SSC. The membrane was placed on a phosphor screen overnight and

visualized on a phosphorimager (Amersham Pharmacia Biotech, Sunnyvale, CA). The probe was removed from the membrane as described [33].

References

1. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL: **Nested retrotransposons in the intergenic regions of the maize genome**. *Science* 1996, **274**:765-8.
2. Li WH, Gu Z, Wang H, Nekrutenko A: **Evolutionary analyses of the human genome**. *Nature* 2001, **409**:847-9.
3. Doolittle WF, Sapienza C: **Selfish genes, the phenotype paradigm and genome evolution**. *Nature* 1980, **284**:601-3.
4. Orgel LE, Crick FH: **Selfish DNA: the ultimate parasite**. *Nature* 1980, **284**:604-7.
5. McDonald JF: **Evolution and consequences of transposable elements**. *Curr Opin Genet Dev* 1993, **3**:855-64.
6. Kidwell MG, Lisch DR: **Perspective: transposable elements, parasitic DNA, and genome evolution**. *Evolution Int J Org Evolution* 2001, **55**:1-24.
7. McDonald JF: **Transposable elements, gene silencing and macroevolution**. *Trends in Ecology and Evolution* 1998, **13**:94-95.
8. Miller WJ, McDonald JF, Nouaud D, Anxolabehere D: **Molecular domestication--more than a sporadic episode in evolution**. *Genetica* 1999, **107**:197-207.

9. Hickey DA: **Selfish DNA: a sexually-transmitted nuclear parasite**. *Genetics* 1982, **101**:519-31.

10. McCollum AM, Ganko EW, Barrass PA, Rodriguez JM, McDonald JF: **Evidence for the adaptive significance of an LTR retrotransposon sequence in a Drosophila heterochromatic gene**. *BMC Evol Biol* 2002, **2**:5.

11. Brosius J: **Genomes were forged by massive bombardments with retroelements and retrosequences**. In: *Transposable Elements and Genome Evolution*. pp. 209-238; 2000: 209-238.

12. Nekrutenko A, Li WH: **Transposable elements are found in a large number of human protein- coding genes**. *Trends Genet* 2001, **17**:619-21.

13. Kumar A, Bennetzen JL: **Plant retrotransposons**. *Annu Rev Genet* 1999, **33**:479-532.

14. Britten RJ: **DNA sequence insertion and evolutionary variation in gene regulation**. *Proc Natl Acad Sci U S A* 1996, **93**:9374-7.

15. Britten RJ: **Mobile elements inserted in the distant past have taken on important functions**. *Gene* 1997, **205**:177-82.

16. Banville D, Boie Y: **Retroviral long terminal repeat is the promoter of the gene encoding the tumor-associated calcium-binding protein oncomodulin in the rat**. *J Mol Biol* 1989, **207**:481-90.

17. Stavenhagen JB, Robins DM: **An ancient provirus has imposed androgen regulation on the adjacent mouse sex-limited protein gene**. *Cell* 1988, **55**:247-54.

18. Mager DL, Hunter DG, Schertzer M, Freeman JD: **Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (HHLA2 and HHLA3)**. *Genomics* 1999, **59**:255-63.

19. Bowen NJ, McDonald JF: **Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside**. *Genome Res* 2001, **11**:1527-40.

20. Ganko EW, Fielman KT, McDonald JF: **Evolutionary History of Cer Elements and Their Impact on the C. elegans Genome**. *Genome Res* 2001, **11**:2066-74.

21. Bowen NJ, McDonald JF: **Genomic analysis of Caenorhabditis elegans reveals ancient families of retroviral-like elements**. *Genome Res* 1999, **9**:924-35.

22. Hodgkin J: **What does a worm want with 20,000 genes?** *Genome Biol* 2001, **2**:COMMENT2008.

23. Kent WJ, Zahler AM: **Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment**. *Genome Res* 2000, **10**:1115-25.

24. McGhee JD, Krause MW: **Transcription Factors and Transcriptional Regulation**. In: *C. Elegans II* Edited by Riddle DL, Blumenthal T, Meyer BJ, Priess JR. pp. 147-184. New York: Cold Spring Harbor Laboratory Press; 1997: 147-184.

25. Kimura M: **The rate of molecular evolution considered from the standpoint of population genetics**. *Proc Natl Acad Sci U S A* 1969, **63**:1181-8.

26. Andrassy I: *A Taxonomic Review of the Suborder Rhabditina (Nematoda: Secernentia)*. Paris: Orstom; 1983.

27. Abdul Kader N, Côté MG: **Isolement, identification et caractérisation de souches québécoises du nématode *Caenorhabditis elegans***. *Fundam. Appl. Nematol.* 1996, **19**:381-389.

28. Egilmez NK, Ebert RH, 2nd, Shmookler Reis RJ: **Strain evolution in Caenorhabditis elegans: transposable elements as markers of interstrain evolutionary history**. *J Mol Evol* 1995, **40**:372-81.

29. Hodgkin J, Doniach T: **Natural variation and copulatory plug formation in Caenorhabditis elegans**. *Genetics* 1997, **146**:149-64.

30. Stenico M, Lloyd AT, Sharp PM: **Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases**. *Nucleic Acids Res* 1994, **22**:2437-46.

31. Thomas WK, Wilson AC: **Mode and tempo of molecular evolution in the nematode caenorhabditis: cytochrome oxidase II and calmodulin sequences**. *Genetics* 1991, **128**:269-79.

32. Wood WB: *The Nematode Caenorhabditis elegans*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 1988.

33. Brenner S: **The genetics of Caenorhabditis elegans**. *Genetics* 1974, **77**:71-94.

34. Chomczynski P, Sacchi N: **Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction**. *Anal Biochem* 1987, **162**:156-9.

35. Sambrook J, Fritsch EF, Maniatis T: *Molecular Cloning A Laboratory Manual*, second edn. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 1989.

Table 2.1: **Presence or absence of retroelement sequences associated with four different genes in strains representing 14 natural populations of *C. elegans*.** (+) indicates presence or detection of the particular element/gene association, (-) indicates absence or no detection of the particular association.

STRAIN	**LOCATION**	***Cer9* / C56G3.2**	***Cer16-2* / 6R55.2**	***Cer2* / F53E10.5**	***Cer16-1* / F20B4.6**
N2	Bristol, England	+	+	+	+
NS3598	Cambridge, England	+	+	+	+
CB4852	Rothamsted, England	+	+	+	+
CB4932	Taunton, England	+	+	+	+
CB4851	Bergerac, France	+	+	+	+
RC301	Freiburg, Germany	+	+	+	+
AB2	Adelaide, Australia	+	-	+	+
CB3191	Altadena, CA	+	+	+	+
DH424	El Prieto Canyon, CA	+	+	+	+
CB4507	Palm Canyon, CA	+	+	+	+
CB4855	Palo Alto, CA	+	+	+	+
TR403	Madison, WI	+	+	-	+
KR314	Vancouver, Canada	-	-	-	+
CB4856	Hawaii	-	+	-	+

Table 2.2: **PCR primers used in this study.** [1]Position is relative to the location of the first nucleotide of the clones 6R55, C56G3, F20B4, or F53E10.

Name	**Sequence**	**Position[1]**
6R55.2 (f)	ACTCCCCAATCACATCTCACTTTC	2508:2531
6R55.2 (r)	AAGATTGAGAGTGAGGTAGTGTGCG	2658:2682
Cer16-2 (f)	TACGACGCTCCGCAATAACG	1211:1230
Cer16-2 (r)	ACGAACCCACAATCACCATCCG	1367:1388
C56G3.2 (f)	GGAGCACTTCGTGATTCTCTCAAG	2352:2375
C56G3.2 (r)	CAAACTGACGCCAGATGTCTTCTAC	2463:2487
Cer9 (f)	CTCCCCCTTTCTCTAACTTAACGC	1803:1826
Cer9 (r)	GGAAGAAGAGTCTAAGAGAGAACTGGC	2238:2264
F20B4.6 (f)	TCAAGAACAGAACGCCTCGTCG	6888:6909
F20B4.6 (r)	AAGGGTTGGGTTTGGTTGGAC	8338:8358
F20B4.6 (f2)	GGTGTGGTTTTTGTGAGTGC	7009:7028
F20B4.6 (f3)	TCGCCTTGAGTATTTCAGTATGGG	8948:8973
Cer16-1 (f)	CCGAGTGACAAACAGCGTATTACAG	7477:7501
Cer16-1 (r)	TCCCGTTATTCCGAAGCGTC	7773:7792
F53E10.5 (f)	CCTCCGTGATTTCATTTATTCGCC	4346:4369
F53E10.5 (r)	GCCGATTTCCGTTCCTTTGTATC	5623:5645
Cer2 (f)	GCGATAGCGTTCTGCTCTTGTG	3749:3770
Cer2 (r)	CCAAACCCCCCAGTGATAGAATAG	3894:3917

Figure 2.1: ***Cer* element/gene associations examined in this study.** Red boxes represent *Cer* elements. Green arrows represent Wormbase predicted genes. Yellow arrows represent ESTs that correspond to the genes. Blue blocks represent predicted exons. The sequence line corresponds to the genomic clone sequence (6R55, C56G3, F20B4, F53E10). Black arrows below the sequence line show the positions of PCR primers used. (A) The 5' LTR of a full-length *Cer16-2* element is contained within the 5' end of 6R55.2, a putative gene with unknown function. (B) The 3' LTR of a fragmented *Cer9* element is positioned in the 5' end of C56G3.2, an aldo/keto reductase homolog. (C) A *Cer16-1* solo LTR is positioned entirely within an intron of gene F20B4.6, a member of the glucosyltransferase family. (D) A *Cer2* solo LTR is contained within the 3' region of F53E10.5, a gene with homology to a helicase domain.

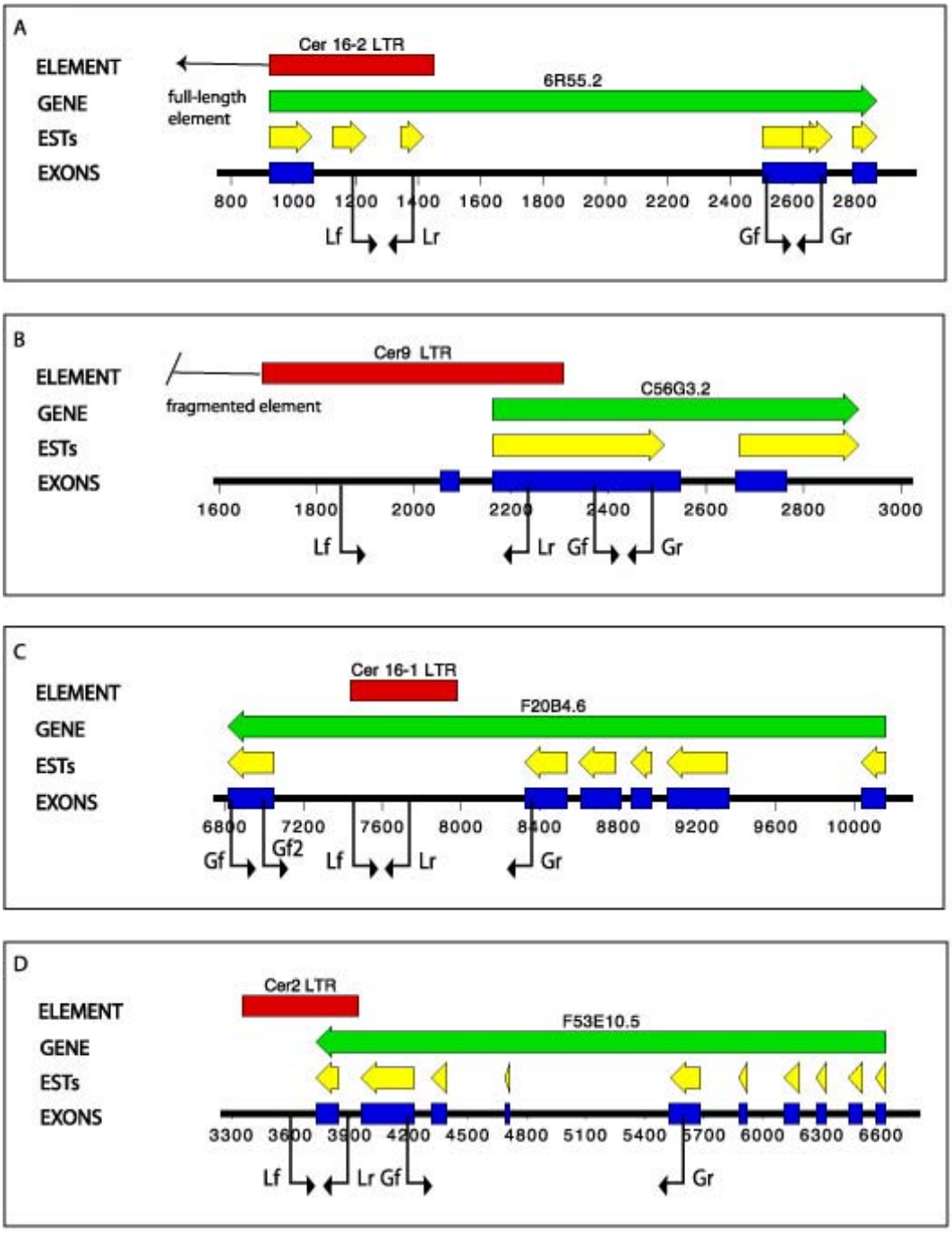
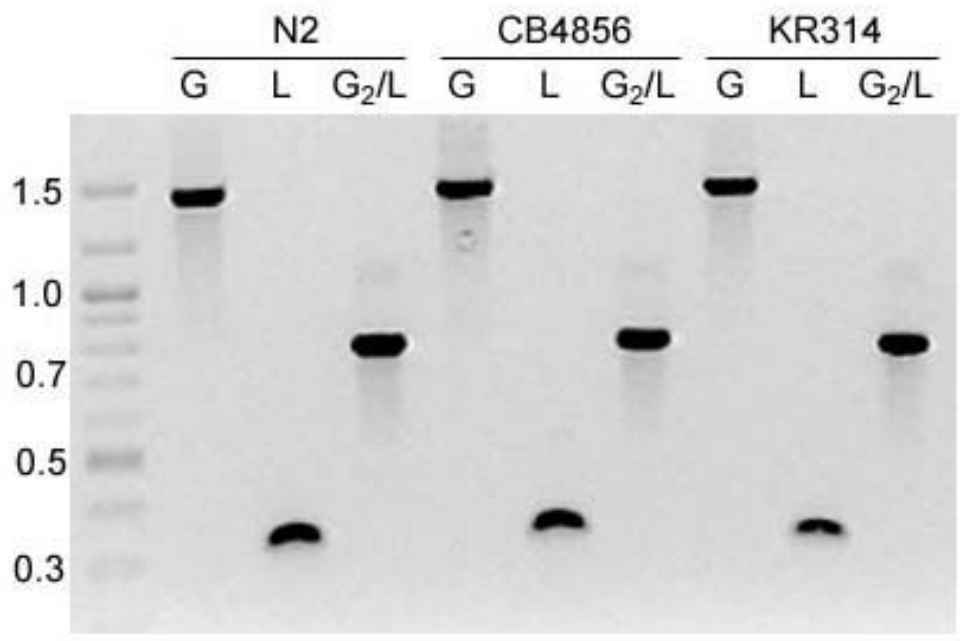

A
ELEMENT
GENE
ESTs
EXONS
Cer 16-2 LTR
full-length element
6R55.2
800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800
Lf Lr Gf Gr
B
ELEMENT
GENE
ESTs
EXONS
Cer9 LTR
fragmented element
C56G3.2
1600 1800 2000 2200 2400 2600 2800 3000
Lf Lr Gf Gr
C
ELEMENT
GENE
ESTs
EXONS
Cer 16-1 LTR
F20B4.6
6800 7200 7600 8000 8400 8800 9200 9600 10000
Gf Gf2 Lf Lr Gr
D
ELEMENT
GENE
ESTs
EXONS
Cer2 LTR
F53E10.5
3300 3600 3900 4200 4500 4800 5100 5400 5700 6000 6300 6600
Lf Lr Gf Gr

Figure 2.2: **PCR analysis to detect the *Cer16-1* LTR in gene F20B4.6 across three representative populations of *C. elegans*.** The *Cer16*-1 LTR is present in all populations tested. A negative image is presented for clarity. Three PCR reactions were performed per strain, per gene. G = product for the gene using gene specific primers, L = product for the LTR using LTR-specific primers, and G2/L = product for the insertion using a second forward gene primer and the reverse LTR primer.


N2
CB4856
KR314
G
L
G2/L
G
L
G2/L
G
L
G2/L
1.5
1.0
0.7
0.5
0.3

CHAPTER 3

EVIDENCE FOR THE ADAPTIVE SIGNIFICANCE OF AN LTR RETROTRANSPOSON SEQUENCE IN A *DROSOPHILA* HETEROCHROMATIC GENE[1]

[1] McCollum, A.M., E.W. Ganko, P.A. Barrass, J.M. Rodriguez, and J.F. McDonald. 2002. *BMC Evolutionary Biology.* 2:5.
Reprinted here with permission of publisher.

Abstract

Background

The potential adaptive significance of transposable elements (TEs) to the host genomes in which they reside is a topic that has been hotly debated by molecular evolutionists for more than two decades. Recent genomic analyses have demonstrated that TE fragments are associated with functional genes in plants and animals. These findings suggest that TEs may contribute significantly to gene evolution.

Results

We have analyzed two transposable elements associated with genes in the sequenced *Drosophila melanogaster y; cn bw sp* strain. A fragment of the *Antonia* long terminal repeat (LTR) retrotransposon is present in the intron of *Chitinase 3* (*Cht3*), a gene located within the constitutive heterochromatin of chromosome 2L. Within the euchromatin of chromosome 2R a full-length *Burdock* LTR retrotransposon is located immediately 3' to *cathD*, a gene encoding cathepsin D. We tested for the presence of these two TE/gene associations in strains representing 12 geographically diverse populations of *D. melanogaster*. While the *cathD* insertion variant was detected only in the sequenced *y; cn bw sp* strain, the insertion variant present in the heterochromatic *Cht3* gene was found to be fixed throughout twelve *D. melanogaster* populations and in a *D. mauritiana* strain suggesting that it maybe of adaptive significance. To further test this hypothesis, we sequenced a 685bp region spanning the LTR fragment in the intron of *Cht3* in strains representative of the two sibling species *D. melanogaster* and *D. mauritiana* (~2.7 million years divergent). The level of sequence divergence between the

two species within this region was significantly lower than expected from the neutral substitution rate and lower than the divergence observed between a randomly selected intron of the *Drosophila Alcohol dehydrogenase* gene (*Adh*).

Conclusions

Our results suggest that a 359 bp fragment of an *Antonia* retrotransposon (complete LTR is 659 bp) located within the intron of the *Drosophila melanogaster Cht3* gene is of adaptive evolutionary significance. Our results are consistent with previous suggestions that the presence of TEs in constitutive heterochromatin may be of significance to the expression of heterochromatic genes.

Background

The potential adaptive significance of transposable elements (TEs) to the host genomes in which they reside is a topic that has been hotly debated by molecular evolutionists for more than two decades. While the biological importance of TEs seemed self-evident to those scientists involved in their initial discovery [e.g., 1, 2], the subsequent realization that TEs could be maintained in populations even while imparting slight selective disadvantage to their hosts [e.g., 3-5] drew into question the presumption of adaptive significance. However, even if TEs can be maintained in populations on a day-to-day basis without providing selective advantage, it does not preclude the possibility that the insertion of TEs in or near genes may, in some instances, be of adaptive advantage.

If TE insertion variants have contributed to adaptive gene evolution, such variants might be expected to be in high frequency or fixed in populations and species. Initial surveys of natural populations of *Drosophila melanogaster* showing that TE insertion alleles are in uniformly low frequency seemed to negate the adaptive hypothesis [6]. However, the sporadic discovery of degenerate TEs or TE fragments as critical components of functional genes in both plants and animals was sufficient to keep the adaptive hypothesis alive throughout the pre-genomic era [7-11].

The current availability of the complete or nearly complete sequence of select genomes representing a variety of species is providing an unprecedented opportunity to examine the frequency and distribution of TEs in eukaryotic genomes. The results have been dramatic. TEs not only comprise a significant fraction of nearly all eukaryotic

genomes thus far sequenced, they have been found to be components of the regulatory and/or coding regions of a surprisingly large number of genes [e.g., 12]. For example, a recent genomic analysis of 13,799 human genes revealed that approximately 4% harbored retrotransposon sequences within protein-coding regions [13]. Similar results have been recently reported for the nematode *Caenorhabditis elegans* [14]. Here we analyze the polymorphism of two LTR retrotransposon / host gene associations across geographically widespread *D. melanogaster* populations and a representative population of the *D. melanogaster* sibling species, *Drosophila mauritiana*.

Results

We have initiated a genomic analysis of LTR retrotransposons present in the *Drosophila melanogaster* genome [e.g., 15]. Of particular interest is identification of genes harboring TEs and determining if these insertion alleles are in high frequency or fixed among natural populations as would be expected from the adaptive hypothesis. We report here the results of an analysis of two LTR retrotransposon-containing genes located on the second chromosome of the sequenced *D. melanogaster y; cn bw sp* strain. These two genes present an interesting contrast in that one of them, *Chitinase 3* (*Cht3)*, is located within constitutive heterochromatin (Genbank accession: AE002743) while the other, *cathD*, is located in a euchromatic region of the chromosome (Genbank accession: AE003839). Our findings demonstrate that while the euchromatic *cathD* insertion variant was not detected in any of the natural populations examined, the insertion variant present in the heterochromatic *Cht3* gene was found to be apparently fixed throughout the

species. These results are consistent with the view that the presence of TEs in constitutive heterochromatin may have relevance to the expression of heterochromatic genes [e.g., 16, 17].

Genomic analysis of the sequenced *y; cn bw sp* strain of *Drosophila melanogaster* identified a full-length *Burdock* LTR retrotransposon located just 3' to the *cathD* gene and a 359bp LTR fragment (complete LTR is 659 bp) of an *Antonia* LTR retrotransposon [15] located within an intron of the *Cht3* gene (Figure 3.1). A set of PCR primers were designed to amplify regions of both genes and retrotransposon sequences. Appropriate pairs of gene and element primers were used to detect the presence or absence of the respective retrotransposon inserts associated with each gene in strains representing 12 geographically dispersed populations of *D. melanogaster*. The results presented in Figure 3.1 and Table 3.1 demonstrate that while the *Burdock* insertion located just 3' to *cathD* gene is not present in any of 12 strains representing a geographically diverse sampling of natural populations, the *Antonia* LTR fragment located in the intron of the heterochromatic *Cht3* gene is fixed in all 12 strains tested.

It is formally possible that the presence of the *Antonia* LTR within the *Cht3* intron was the result of a chance fixation event prior to the expansion of *D. melanogaster* around the world. Thus, to further test the adaptive hypothesis we compared the level of sequence divergence within the LTR and its flanking intronic sequence between the two sibling species *Drosophila melanogaster* and *Drosophila mauritiana*. If the LTR-containing intron is under stabilizing selection, a lower than neutral rate of substitution would be expected. A total of 685bp of the *Cht3* intron was sequenced. This region spans 264bp of the 359 bp *Antonia* LTR fragment. The sequence of this region in a *D.*

melanogaster (Dimonika, Africa) and *D. mauritiana* (Mauritius, Africa) strain was aligned with the homologous region in the sequenced *D. melanogaster y; cn bw sp* strain. The two *melanogaster* strains were 100% identical. The *melanogaster* sequences were found to be only 1.3% (9 substitutions/685 nucleotide sites) diverged from that of *D. mauritiana.* This value is significantly less than half of the expected 4.3 % (±2.7) divergence based on the *Drosophila* neutral substitution rate of 0.016 (±0.005) substitutions/site/million year [18] over the estimated 2.7 million years separating the two species [19].

To directly compare the substitution rate for the *Cht3* intron with that of another *Drosophila* gene intron, we randomly selected intron 1 of the *Drosophila* alcohol dehydrogenase (*Adh*) gene. *Adh* is a widely studied *Drosophila* gene and it has been sequenced in several *Drosophila* species including *D. melanogaster*, accession X60793 [20] and *D. mauritiana,* accession M19264 [21]. The sequence divergence between *D. melanogaster* and *D. mauritiana* in the *Adh* intron 1 (7.9%, Figure 3.4), is higher than that for the LTR containing *Cht3* intron (1.3%). These results strongly suggest that conservative selection has been operating on the LTR containing intron associated with the *Drosophila Cht3* gene over the past 2.7 million years.

Discussion

For many years, constitutive heterochromatin was considered to be of little or no functional significance [22]. This view seemed to be supported by early molecular studies showing that heterochromatin consists almost exclusively of highly repeated and

middle repetitive DNA [e.g., 23, 24]. The middle repetitive fraction was viewed as the descendent of once active TEs that had the misfortune of inserting into transcriptionally inert heterochromatin at some point in their evolutionary history [e.g., 6, 20]. The view of heterochromatin as a genetic wasteland gradually changed with the mapping of a number of functionally important *Drosophila* genes to constitutive heterochromatin [e.g., 24-31]. Reexamination of *Drosophila* constitutive heterochromatin revealed that long stretches of highly repetitive DNA are interrupted by "islands" of retrotransposon sequences [e.g., 32, 33]. *Drosophila* genes in heterochromatin are typically associated with these islands of retrotransposons [2, 31, 34-36]. It has been suggested that transposable elements inserted into heterochromatin may locally alter chromatin structure [e.g., 16]. Our results suggest that in at least some instances, the association of heterochromatic genes with transposable element sequences may be of adaptive significance.

Conclusions

The results presented here are consistent with the hypothesis that a 359 bp fragment of the *Antonia* retrotransposon located within the intron of the heterochromatic *Drosophila melanogaster Cht3* gene may be of adaptive evolutionary significance. Further genomic and molecular analyses will be required to assess the general importance of LTR retrotransposon sequences to the evolution of heterochromatic gene structure and function.

Materials and Methods

Gene Region Annotation

BLASTS of sequenced DNA turned up several instances of genes proximal to an LTR retrotransposon. Sequence retrieval was initiated via BLASTN searches (default parameters- [37]) against the BDGP (http://www.fruitfly.org) and GenBank (http://www.ncbi.nlm.nih.gov) databases using LTRs from previously identified *Drosophila* retroelements as queries [15]. Results with E-values < e^{-10} were annotated on the corresponding clone, whereupon visual inspection of several annotations confirmed the presence of retroelements proximal to known genes. Selected genes were BLASTed against NCBI's EST database and mapped along with predicted transcript structures from Flybase (http://www.flybase.org). Chromosomal location of clones was also determined from Flybase.

PCR

D. melanogaster strains from Dimonika, Niamey, Swaziland, Kenia, Capetown, Cotonake, and India were obtained from Charles F. Aquadro, Cornell University. Germany, Italy, and Antilles strains were obtained from Nikolaj Junakovic, Universitá la Sapienza, Rome, Italy. California and Athens strains are from Daniel Promislow, University of Georgia. *D. melanogaster y;cn bw sp* strain was obtained from the Bloomington, IN, stock center. The *D. mauritiana* (241.0) strain was provided by the Bowling Green, OH, *Drosophila* stock center.

PCR primers were designed with MacVector 7.0 (http://www.gcg.com) and synthesized by Integrated DNA Technologies (Coralville, IA) (Table 3.2). Three PCR reactions were performed per strain, per gene. For all PCR reactions, 1.0µl of a single fly DNA prep [38] was used and amplification was performed in a Hot Top equipped Robocyler Gradient 96 (Stratagene, La Jolla, CA). 10µl of product was separated on a 1% agarose gel in 0.5x TBE running buffer containing 0.25µg mL^{-1} ethidium bromide. Gel images were visualized by UV transillumination.

<u>*Cht3* PCR</u> - The PCR products for primer set *cht3(f)* and *cht3(r)* and primer set *Antonia LTR(f)* and *Antonia LTR(r)* were amplified in a 25µl reaction containing 3mM $MgCl_2$, 10X PCR buffer supplied by Pierce (Rockford, IL), 5% DMSO, 0.2mM dNTPs, 0.5µM of each primer, and 0.5U of Taq DNA polymerase supplied by Pierce [Rockford, IL]. The program consisted of an initial incubation at 94°C for 3 min for 1 cycle, a 30 cycle extension at 94°C for 30 sec, 56°C for *cht3(f)*/*cht3(r)* primer set or 57°C for *Antonia LTR(f)*/*Antonia LTR(r)* primer set for 30 sec, 72°C for 1 min 30 sec, and a 1 cycle final extension of 72°C for 5 min. The PCR products for primer set *cht3(f2)* and *LTR(r)* were amplified in a 25µl reaction containing Expand Long Template PCR System 10X PCR buffer #1 supplied by Roche (Indianapolis, IN), 0.35mM dNTPs, 0.32µM of each primer, and 1.3U of Expand Long Template PCR System DNA polymerase mix supplied by Roche (Indianapolis, IN). The program consisted of an initial incubation at 94°C for 3 min for 1 cycle, a 30 cycle extension at 94°C for 30 sec, 52°C for 30 sec, 68°C for 3 min, and a 1 cycle final extension at 68°C for 5 min.

<u>*cathD* PCR</u> - The reaction mix and program used for all sets of primers are the same as those described for primer set *cht3(f)* and *cht3(r)* and primer set *Antonia LTR(f)*

and *Antonia LTR(r)* in the *Cht3* PCR (above). The annealing temperature for primer set *cathD(f)* and *cathD(r)* is 58°C, for primer set *Burdock LTR(f)* and *Burdock element(r)* is 59°C, and for primer set *cathD(f)* and *Burdock element(r)* is 56°C.

Sequencing

PCR products of the *Cht3* intron were sequenced in the Molecular Genetics Instrumentation Facility at the University of Georgia. Sequences were aligned with MacVector 7.0 and compared to the published *y; cn bw sp* strain. Substitutions and insertion/deletion sites (indels) were summed for each sequence product and compared to the expected divergence based upon the neutral substitution rate. The expected number of polymorphisms between *D. melanogaster* and *D. mauritiana* was calculated based on the *Drosophila* neutral substitution rate of .016 (±0.005) substitutions per site/million years [18] on 685bp over a divergence time of 2.7 million years [19].

Acknowledgements

E.W.G supported through an NIH Genetics Training Grant. This work supported by a National Institutes of Health grant to J.F.M.

References

1. McClintock B: **Chromosome organization and genic expression**. *Cold Spr. Harb. Symp. Quant. Biol.* 1951, **16**:13-17.

2. Shapiro J: **DNA insertion elements and the evolution of chromosome primary structure**. *Trends Biochem. Sci.* 1977, **2**:622-627.

3. Doolittle WF, Sapienza C: **Selfish genes, the phenotype paradigm and genome evolution**. *Nature* 1980, **284**:601-603.

4. Hickey DA: **Selfish DNA: a sexually-transmitted nuclear parasite**. *Genetics* 1982, **101**:519-531.

5. Orgel LE, Crick FH: **Selfish DNA: the ultimate parasite**. *Nature* 1980, **284**:604-607.

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

7. Britten RJ: **DNA sequence insertion and evolutionary variation in gene regulation**. *Proc Natl Acad Sci U S A* 1996, **93**:9374-9377.

8. Britten RJ: **Mobile elements inserted in the distant past have taken on important functions**. *Gene* 1997, **205**:177-182.

9. Brosius J: **Retroposons--seeds of evolution**. *Science* 1991, **251**:753.

10. Makalowski W, Mitchell GA, Labuda D: **Alu sequences in the coding regions of mRNA: a source of protein variability**. *Trends Genet* 1994, **10**:188-193.

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

12. Brosius J: **Genomes were forged by massive bombardments with retroelements and retrosequences**. In: *Transposable Elements and Genome Evolution* Edited by McDonald JF. pp. 209-238. Dordrecht: Kluwer Academic Publishers; 2000: 209-238.

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

14. Ganko EW, Fielman KT, McDonald JF: **Evolutionary History of Cer Elements and Their Impact on the C. elegans Genome**. *Genome Res* 2001, **11**:2066-2074.

15. Bowen NJ, McDonald JF: **Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside**. *Genome Res* 2001, **11**:1527-1540.

16. Gatti M, Pimpinelli S: **Functional elements in Drosophila melanogaster heterochromatin**. *Annu Rev Genet* 1992, **26**:239-275.

17. Spradling AC: **Transposable elements and the evolution of heterochromatin**. *Soc Gen Physiol Ser* 1994, **49**:69-83.

18. Li WH: *Molecular Evolution*. Sunderland, MA: Sinauer; 1997.

19. Li YJ, Satta Y, Takahata N: **Paleo-demography of the Drosophila melanogaster subgroup: application of the maximum likelihood method**. *Genes Genet Syst* 1999, **74**:117-127.

20. Laurie CC, Bridgham JT, Choudhary M: **Associations between DNA sequence variation and variation in expression of the Adh gene in natural populations of Drosophila melanogaster**. *Genetics* 1991, **129**:489-499.

21. Cohn VH, Moore GP: **Organization and evolution of the alcohol dehydrogenase gene in Drosophila**. *Mol Biol Evol* 1988, **5**:154-166.

22. Hannah A: **Localization and function of heterochromatin in *Drosophila melanogaster***. *Adv. Genet.* 1951, **4**:87-125.

23. Appels R, Peacock WJ: **The arrangement and evolution of highly repeated (satellite) DNA sequences with special reference to Drosophila**. *Int Rev Cytol Suppl* 1978, **Suppl**:69-126.

24. Brutlag DL: **Molecular arrangement and evolution of heterochromatic DNA**. *Annu Rev Genet* 1980, **14**:121-144.

25. Hilliker AJ, Holm DG: **Genetic analysis of the proximal region of chromosome 2 of Drosophila melanogaster. I. Detachment products of compound autosomes**. *Genetics* 1975, **81**:705-721.

26. Hilliker AJ: **Genetic analysis of the centromeric heterochromatin of chromosome 2 of Drosophila melanogaster: deficiency mapping of EMS-induced lethal complementation groups**. *Genetics* 1976, **83**:765-782.

27. Ganetzky B: **On the components of segregation distortion in Drosophila melanogaster**. *Genetics* 1977, **86**:321-355.

28. Pimpinelli S, Sullivan W, Prout M, Sandler L: **On biological functions mapping to the heterochromatin of Drosophila melanogaster**. *Genetics* 1985, **109**:701-724.

29. Marchant GE: **Genetic analysis of the heterochromatin of chromosome 3 in *Drosophila melanogaster*. I. Products of compound autosome detachment.** *Genetics* 1988a, **120**:503-517.

30. Marchant GE: **Genetic analysis of the heterochromatin of chromosome 3 in *Drosophila melanogaster*. II. Vital loci identified through EMS mutagenesis.** *Genetics* 1988b, **120**:519-532.

31. Devlin RH, Holm DG, Morin KR, Honda BM: **Identifying a single-copy DNA sequence associated with the expression of a heterochromatic gene, the light locus of Drosophila melanogaster**. *Genome* 1990, **33**:405-415.

32. Caizzi R, Caggese C, Pimpinelli S: **Bari-1, a new transposon-like family in Drosophila melanogaster with a unique heterochromatic organization**. *Genetics* 1993, **133**:335-345.

33. Pimpinelli S, Berloco M, Fanti L, Dimitri P, Bonaccorsi S, Marchetti E, Caizzi R, Caggese C, Gatti M: **Transposable elements are stable structural components of Drosophila melanogaster heterochromatin**. *Proc Natl Acad Sci U S A* 1995, **92**:3804-3808.

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

35. Berghella L, Dimitri P: **The heterochromatic rolled gene of Drosophila melanogaster is extensively polytenized and transcriptionally active in the salivary gland chromocenter**. *Genetics* 1996, **144**:117-125.

36. Dimitri P, Junokovic, N., and B. Arca: **Nested transposons within the intron II of rolled, a heterochromatic gene of *Drosophila melanogaster***. *Dros Res Conf* 1999, **40**:663B.

37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**:3389-3402.

38. Gloor GB, Preston CR, Johnson-Schlitz DM, Nassif NA, Phillis RW, Benz WK, Robertson HM, Engels WR: **Type I repressors of P element mobility**. *Genetics* 1993, **135**:81-95.

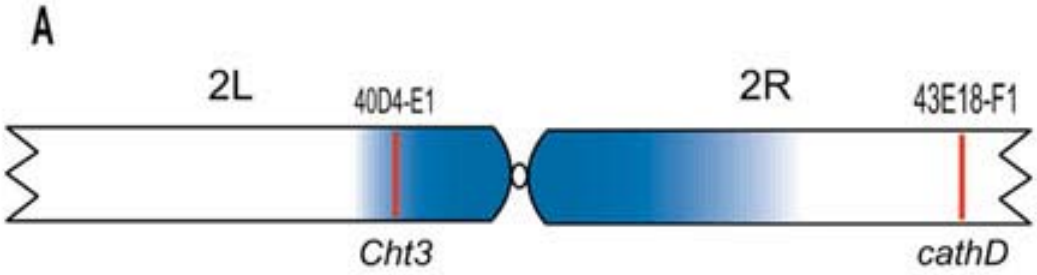
Table 3.1: **Presence or absence of retroelement sequence associated with *cathD* and *Cht3* genes in strains representing 12 natural populations of *D. melanogaster***. Males and females from each strain were tested. (+) indicates presence of retroelement sequence, (-) indicates absence of retroelement sequence.

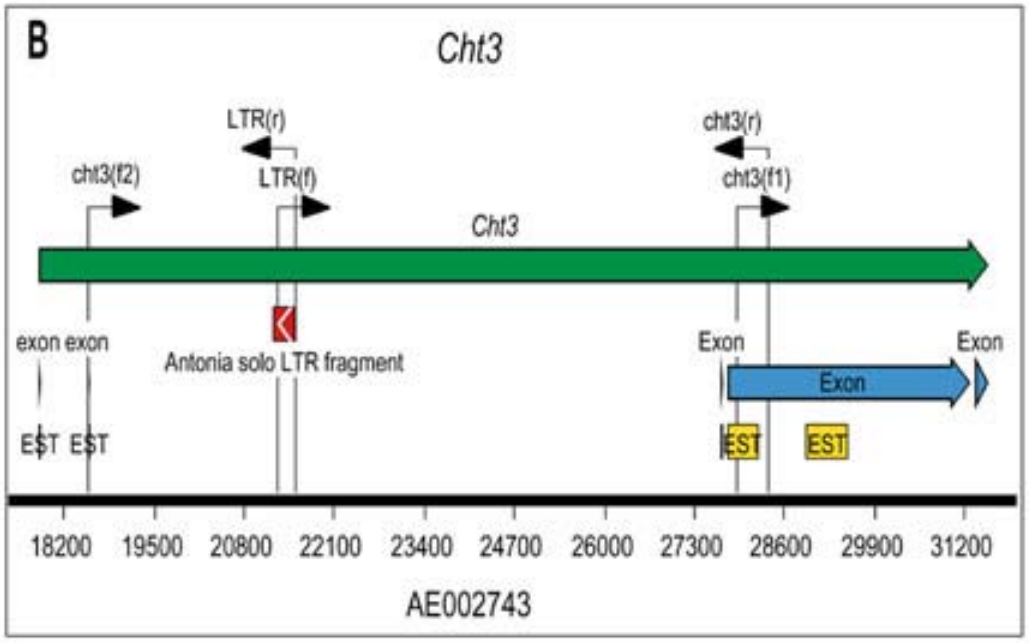
Geographic area	Strain	*cathD / Burdock*	*Cht3 / Antonia*
Lab stock	*y; cn bw sp*	+	+
Americas	Athens	-	+
	California	-	+
	Antilles	-	+
Europe	Germany	-	+
	Italy	-	+
Africa	Capetown	-	+
	Cotonake	-	+
	Dimonika	-	+
	Kenia	-	+
	Niamey	-	+
	Swaziland	-	+
Asia	India	-	+

Table 3.2: **Primers used for PCR analysis**. [1] Position is relative to location of first nucleotide within clone, AE002743 for *cht3* or AE003839 for *cathD*.

Primer	Sequence	Position1
cht3(f)	5'-TGATGCCATACTCCTACTCCGTA	27910:27934
cht3(f2)	5'-ATGAAAAACGGATGGACAGCC-3	18549:18569
cht3(r)	5'-CATTCCTGTTTGCCAACCCC-3'	28395:28376
Antonia LTR(f)	5'-TTAAGCGAACGTCGGAGAC-3'	21299:21317
Antonia LTR(r)	5'-CCACTAGAAGGGTGAAAACTGC	21570:21549
cathD(f)	5'-GGTGAAGCCGCCATTTTACG-3'	220780:220799
cathD(r)	5'-CGCCCAGCACAAACTTGATTAC-	221240:221219
Burdock LTR(f)	5'-TGACCGACGCTTCTAATCTTCC-3	221638:221659
Burdock element(r)	5'-GGTTGGCAGTATGGGAACTTAG	221918:221895

Figure 3.1: **Genomic structure of the *Cht3* and *cathD* genes in the *Drosophila melanogaster* genome**. **(A)** Chromosome 2 illustrating location of *Cht3* and *cathD* genes (red lines) in reference to constitutive heterochromatin (in blue) [34]. Numbers above each red line refer to Flybase cytogenetic placement. (Chromosome not drawn to scale). **(B & C)** Green arrows represent Flybase-predicted gene regions with corresponding identification. Yellow blocks depict ESTs concordant to the predicted gene region. Blue boxes are predicted exon regions. Red boxes denote LTR position and internal arrows indicate orientation of retroelement. The black line and numbers represent position along the genomic clone sequence which is identified below the figure. Black arrows indicate direction and location of forward (f) or reverse(r) PCR primers. **(B)** An *Antonia* LTR fragment (359nt) is inserted in an intron of*Cht3* in 12 geographically distinct *Drosophila melanogaster* strains. **(C)** A full-length *Burdock* retroelement, only present in the sequenced *y; cn bw sp* strain, overlaps the predicted exon boundaries of the *cathD* gene by 6nt.


A
2L
40D4-E1
2R
43E18-F1
Cht3
cathD


B
Cht3
LTR(r)
cht3(r)
cht3(f2)
LTR(f)
cht3(f1)
Cht3
exon exon
Antonia solo LTR fragment
Exon
Exon
Exon
EST EST
EST
EST
18200 19500 20800 22100 23400 24700 26000 27300 28600 29900 31200
AE002743

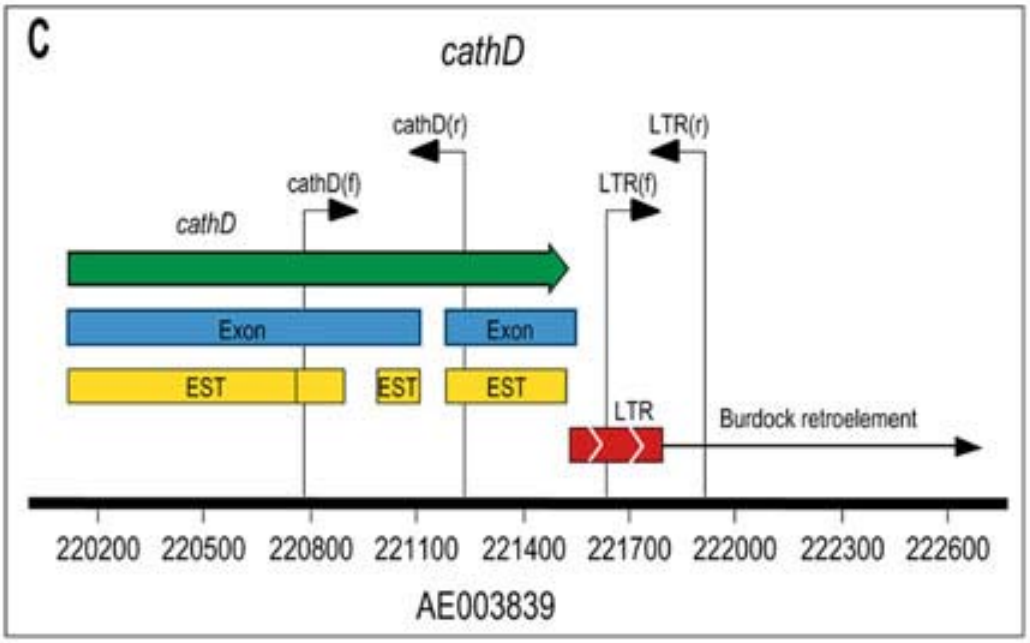

C
cathD
cathD(r)
LTR(r)
cathD(f)
LTR(f)
cathD
Exon
Exon
EST
EST
EST
LTR
Burdock retroelement
220200 220500 220800 221100 221400 221700 222000 222300 222600
AE003839

Figure 3.2**: PCR analysis testing for the presence of an LTR retroelement feature in two genes, *Cht3* and *cathD,* across three representative *Drosophila* strains**. A negative image is presented for visual clarity. Three PCR reactions were performed per strain, per gene. M = 1 kb ladder, M2 = 0.1 kb ladder. **(A)** An *Antonia* LTR fragment is fixed in the intron of the heterochromatic *Cht3* gene in all 12 tested strains (only three shown). *Cht3* - G= *cht3* primers (f+r), expected product= 488bp. L= *Antonia* LTR primers (f+r), expected product= 272bp. G_2/L = *cht3*(f2) + *Antonia* LTR (r) primers, expected product= 3022bp. **(B)** A full-length Burdock LTR retrotransposon is found to be associated with cathD only in the sequenced *y; cn bw sp* strain. *cathD* - G= *cathD* primers (f+r), expected product= 461bp. L= *Burdock* primers (f+r), expected product= 280bp. G/L = cathD(f) and *Burdock* element (r), expected product= 1139bp.

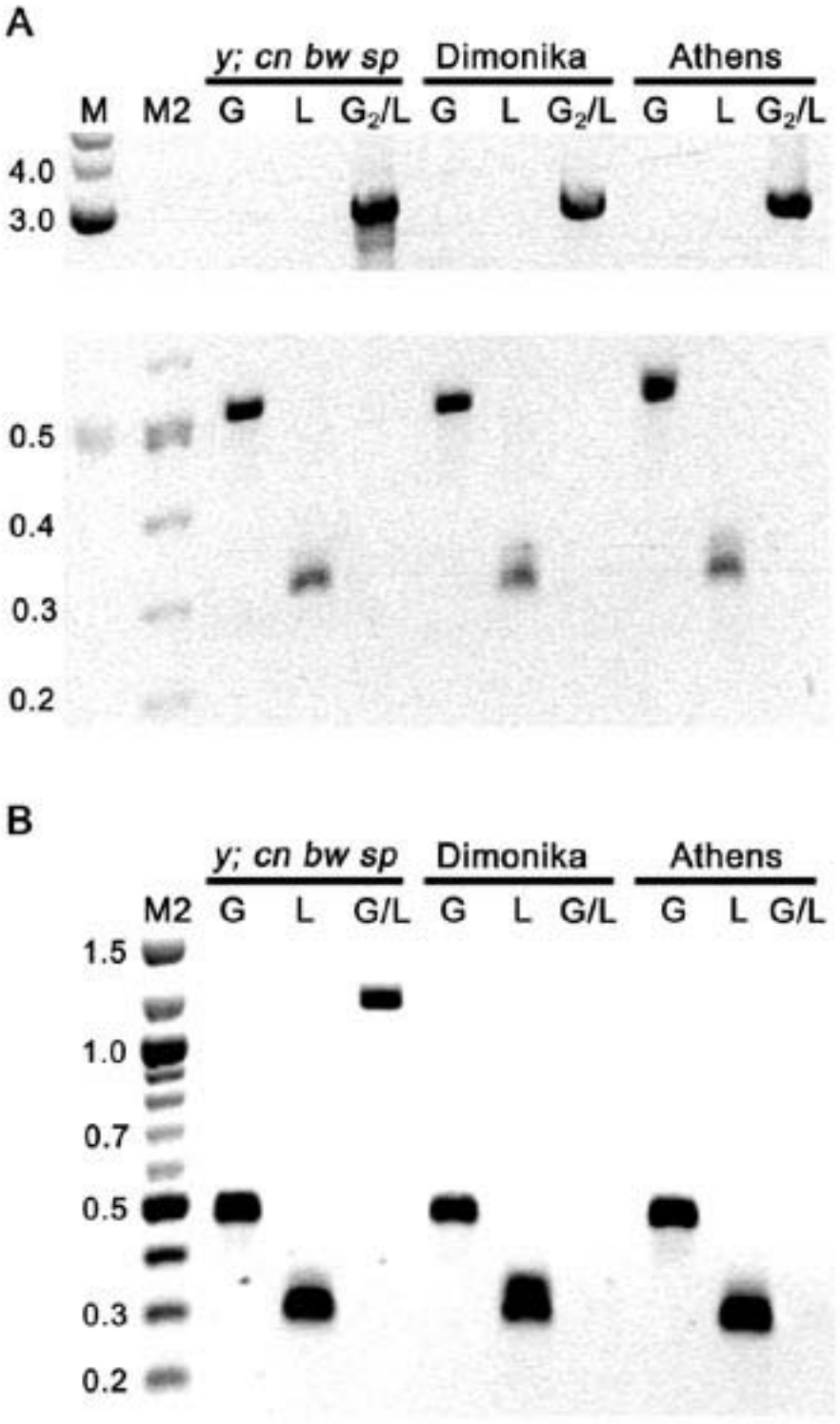

A
y; cn bw sp
Dimonika
Athens
M
M2
G
L
G2/L
G
L
G2/L
G
L
G2/L
4.0
3.0
0.5
0.4
0.3
0.2
B
y; cn bw sp
Dimonika
Athens
M2
G
L
G/L
G
L
G/L
G
L
G/L
1.5
1.0
0.7
0.5
0.3
0.2

Figure 3.3: **Nucleotide alignment of a 685 bp Cht3 intron fragment in *D.melanogaster* and *D.mauritiana***. Cht3 intron sequence from the *Drosophila melanogaster y; cn bw sp* strain (accession AE002743). The Antonia LTR stretches from bp 1 – 264, where a black diamond (♦) indicates the end of LTR sequence. Strains representing the *D. melanogaster*, Africa (Dimonika) population and a strain representing the *D. mauritiana,* Mauritius population were sequenced. Sequences were aligned using MacVector (See Materials and Methods for details).

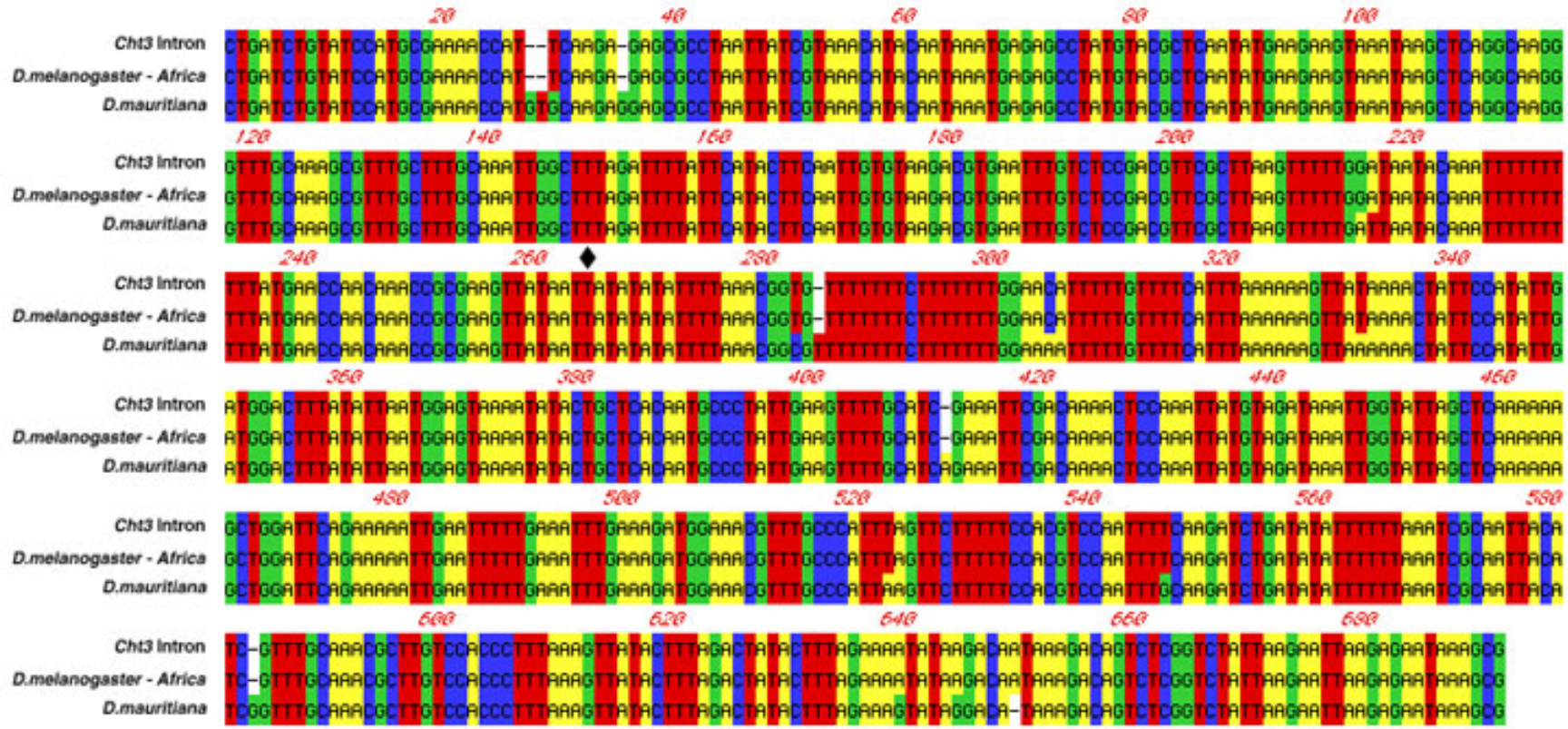
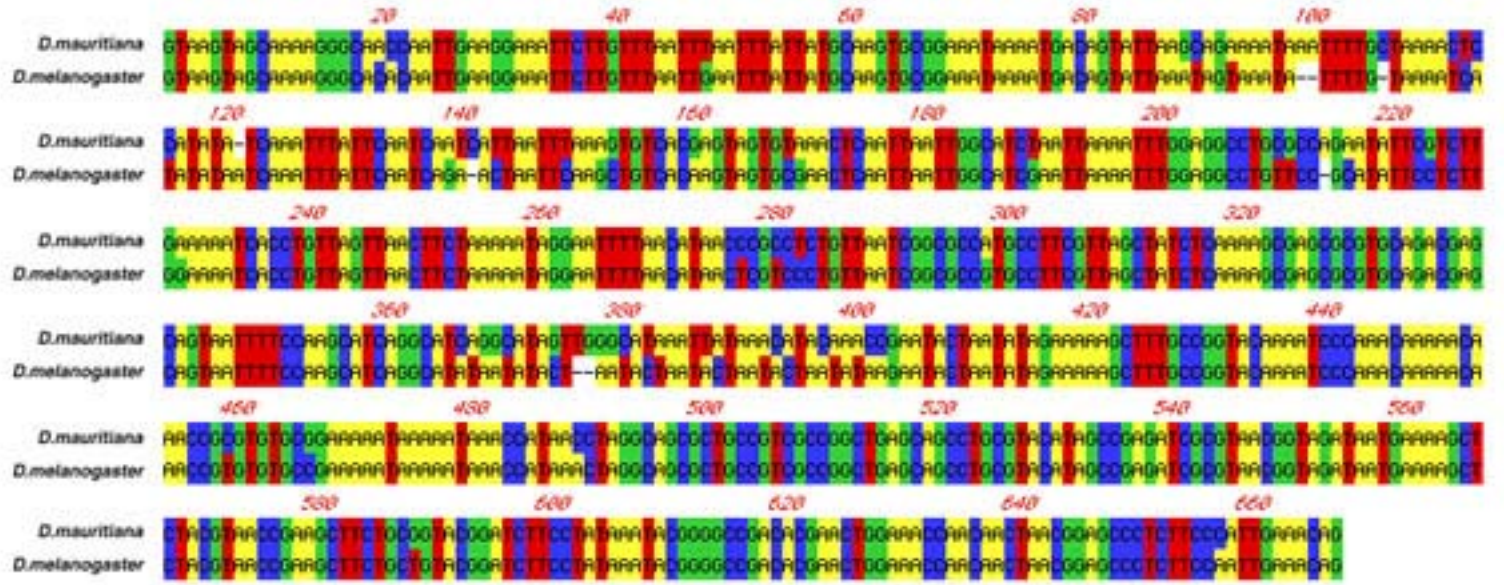

Cht3 Intron
D.melanogaster - Africa
D.mauritiana

Figure 3.4: **Nucleotide alignment of the 659 bp intron 1 of the *Adh* gene in *Drosophila melanogaster* and *Drosophila mauritiana*.** Sequences obtained through GenBank for *D.melanogaster* (accession: X60793, [20] and *D.mauritiana* (accession: M19264, [21]). Sequences were aligned using MacVector (See Materials and Methods for details).


D.mauritiana
D.melanogaster

CHAPTER 4

CONCLUSIONS

Transposable elements (TEs) are ubiquitous components of genomes and their significance in genome evolution is only beginning to be evaluated. This thesis has presented an initiation of experiments in *Caenorhabditis elegans* and *Drosophila. melanogaster* to demonstrate the impact of TEs on host genes in which they reside.

In *C. elegans*, each TE/gene association surveyed is present in high frequency in the species, suggesting that these associations may be of adaptive significance to the host genome. Further molecular tests should be conducted in order to understand the functional significance of these TE insertions on the host genes in which they reside.

In *Drosophila*, it was shown that a heterochromatic TE/gene association is fixed in the species *D. melanogaster*, indicating adaptive significance of this insertion event. Results also show that this insert is under selective pressure in both *D. melanogaster* and sister taxa *D. mauritiana*. These results are consistent with suggestions that the presence of TEs in constitutive heterochromatin may be of significance to the expression of heterochromatic genes.

This thesis has presented a series of experiments that can be used to elucidate the impacts of TEs on genes in which they reside. Complete genome sequences are giving scientists an unprecedented opportunity to study genomes, and molecular approaches to accompany genome sequences give scientists a unique view of genome evolution.