ANALYSIS OF LTR RETROTRANSPOSON CONTRIBUTION TO GENE FUNCTION IN *C. ELEGANS* AND *D. MELANOGASTER*

by

ERIC WALTER GANKO

(Under the Direction of John F. McDonald)

ABSTRACT

Transposons are mobile DNA sequences found in varying abundance within host eukaryotic genomes. One transposon class, the retrotrantransposons, produce RNA copies that are reverse-transcribed into DNA. The DNA copy is then reinserted into the host's genome. To produce transcripts, retrotransposons encode regulatory features including transcriptional promotion and termination signals. Due to their encoded regulatory features and insertional nature, retrotransposons may influence genes in the host organism.

Using the sequenced model organisms *Caenorhabditis elegans* and *Drosophila melanogaster* a number of LTR retrotransposons in close promimity to host genes have been identified. In the worm *C. elegans*, a significantly greater number of transposons are found 500-1000 bp upstream of genes than predicted by random insertion models. In all, 63% of LTR retrotransposon sequences (79/124) are located within 1 kb of a gene or within gene boundaries. Many genes with a nearby LTR retrotransposon lack homology to other species and may be nematode specific. In the genome of the fruitfly *D. melanogaster*, 33.4% of LTR retrotransposon sequences (228/682) are within 1 kb of a gene or within gene boundaries. Genes with an external response function were found to have a neighboring transposon more often than

expected, while genes with metabolic and cell differentiation functions had fewer neighboring transposons. Results in *C. elegans* and *D. melanogaster* are consistent with the hypothesis that LTR retrotransposons may contribute to the structural and/or regulatory evolution of genes.

INDEX WORDS: Genetics, Transposable Element, Molecular Evolution, Long Terminal Repeat (LTR) Retrotransposon, *Drosophila melanogaster*, *Caenorhabditis elegans*

ANALYSIS OF LTR RETROTRANSPOSON CONTRIBUTION TO GENE FUNCTION IN *C. ELEGANS* AND *D. MELANOGASTER*

by

ERIC WALTER GANKO

B.S., Eckerd College, 1998

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2004

ANALYSIS OF LTR RETROTRANSPOSON CONTRIBUTION TO GENE FUNCTION IN *C. ELEGANS* AND *D. MELANOGASTER*

by

ERIC WALTER GANKO

Major Professor: John F. McDonald

Committee: Eileen Kraemer
Jonathan Arnold
Michael McEachern
Ethan W. Taylor

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2004

ACKNOWLEDGEMENTS

Thanks goes to the Genetics Department and the Graduate School for financial support over the last five years. I would like to thank the members of my committee for their patience and support through the graduation process. Paul Schliekelman was very influential with statistical advice on several projects. I am grateful to my advisor, John McDonald, and the other members of the lab for advice and input on my projects. Special thanks to the undergraduates who have contributed: Casey Greene, Judson Lewis, and Vik Bhattacharjee.

Of course, science is only part of my life. I have been blessed with a very supportive family – thanks Mom and Dad - and of course, Leandra.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

INTRODUCTION

Transposable elements (TEs) are DNA sequences with the ability to move from one genomic position to another. First described by Barbara McClintock in the 1940's (McClintock 1946, 1984), thousands of elements have now been identified and classified. There are two major classes of TEs identified by their transposition mechanism (Finnegan 1992). Class II elements are the DNA transposons, which move via a "cut & paste" mechanism. The self-encoded enzyme transposase excises the sequence from the chromosome, followed by reinsertion elsewhere in the genome. Class I elements, also known as retroelements, utilize a "copy & paste" mechanism via an RNA intermediate and the element-encoded enzyme reverse transcriptase (RT, Berg and Howe 1989). In general, an RNA intermediate is transcribed from an existing element by host transcription complexes and converted to DNA by the RT enzyme. After conversion to DNA, the new copy is integrated into the genome. There are several types of class II elements, including long interspersed nuclear elements (LINES) and long terminal repeat (LTR) retrotransposons. Short interspersed nuclear elements (SINES) are non-autonomous retroelements that do not encode enzymes and therefore utilize the enzymes of other autonomous retroelements (Boeke and Stoye 1997).

Transposable elements contribute to host genome size

The abundance of TEs in eukaryotic genomes varies considerably among species. Typically, organisms with small genomes contain relatively few TEs, while organisms with large genomes contain many TEs (Table 1.1). For example, it is estimated that 3% of the 12 Mbp *S. cerevisiae* genome (Kim et al. 1998), 6% of the 100 Mbp *Caenorhabditis elegans* genome (Consortium 1998), and 15% of the 180 Mbp *Drosophila melanogaster* genome (Hoskins et al. 2002; Kaminker et al. 2002) are composed of TEs. In comparison, at least 34% of the 3,000 Mbp mouse genome, 43% of the 3,200 Mbp human genome and 50% of the 4,800 Mbp barley genome are composed of transposons, primarily class I retroelements (Smit 1999; Li et al. 2001; Rostoks et al. 2002; Waterston et al. 2002). It has recently been estimated that the human genome has expanded nearly 20% over the last 50 million years, almost entirely due to new retrotransposon insertions (Liu et al. 2003). Up to 90% of the genomes of some eukaryotes are composed of TEs (i.e. lillies, wheat; Table 1.1). The cumulative transposon load can have direct effects on the evolution of an organism.

Impact of transposable elements on a host

When Barbara McClintock identified transposons approximately 60 years ago, she described a system of "controlling elements" that caused chromosome breakage, especially in response to stress (McClintock 1948, 1984). From these results it was theorized that transposable elements may provide potential adaptive advantages to a host organism in the form of genome restructuring, and that their presence may therefore be maintained over time (McClintock 1951; Shapiro 1977; McClintock 1984). In the early 1980's, an alternate theory emerged when it was shown that transposons could be maintained in a population as neutral or even as deleterious components of the genome (Doolittle and Sapienza 1980; Orgel and Crick

1980; Hickey 1982; Charlesworth and Langley 1989; Charlesworth, Sniegowski and Stephan 1994). According to this view, those TEs located in or near genes are likely to be detrimental to gene function and will be removed by natural selection. Indeed, TEs make up a large percentage of many genomes (Table 1.1), and insertions into genes can create mutations. For example, transposons are a well-known source of spontaneous mutations in Drosophila, potentially accounting for more than half of all mutations (Green 1988). In humans, transposon insertions are involved with Charcot-Marie-Tooth disease (Reiter et al. 1999), Haemophilia A (Kazazian et al. 1988), and at least 20 other diseases including several cancers (e.g. Deininger and Batzer 1999; Ostertag and Kazazian 2001). Since transposons are repeats, they may also act as a strong recombination force (Deininger and Batzer 2002) and can create regions of chromosome instability (Deininger et al. 2003; Jurka et al. 2004).

Other findings in molecular biology and genomics indicate that an entirely selfish view of transposons is shortsighted, leading to the reemergence of an adaptive role for transposons in genomes (McDonald 1993; McDonald 1995; Brosius 1999; Kidwell and Lisch 2001). For example, the evolution of novel genetic function may result from some TE-induced recombination events (Eichler and Sankoff 2003), such that transposons are "likely drivers of evolutionary change" (Deininger et al. 2003). Transposons have been shown to be important contributors to yeast double-strand break repair (Garfinkel 1997), Drosophila telomere maintenance (Pardue et al. 1996), and have also been implicated in mammalian DNA repair (Morrish et al. 2002). On a more direct level, a number of transposons have been shown to significantly affect the expression and coding of individual genes (e.g. Kapitonov and Jurka 1999; Makalowski 2000; Medstrand, Landry and Mager 2001 - see also Table 1.2).

Several studies have taken a broader view by investigating an entire genome for potential transposon–gene interactions. For example, recent analyses in the human genome indicate that TE sequences are found in the intron and exon regions of ~4% of genes (Nekrutenko and Li 2001), in the untranslated regions (UTR) of ~27% of genes (van de Lagemaat et al. 2003), and in ~25% of promoter regions (Jordan et al. 2003). In Arabidopsis plants an estimated 5% of genes have a transposon component (Kumar and Bennetzen 1999). Given their ubiquitous presence, the possibility for mutation and the potential for adaptive benefit, transposons are important genome components that may provide evolutionary plasticity.

LTR retrotransposon biology

Class I retroelements make up a significant portion of the TEs identified in genomes and possess features that may become adopted by a host gene. The retroelements analyzed in the three chapters of this thesis are LTR retrotransposons, which have a life cycle analogous to that of infectious retroviruses (Boeke et al. 1985). A model full-length LTR retrotransposon features genes (*gag*, *pol* and, in some cases, *env*) flanked by long terminal repeats (LTRs) (Boeke and Stoye 1997). The *gag* coding region encodes proteins that form the physical structure of the retroviral-like virion (Coffin, Hughes and Varmus 1997). The *pol* coding region encodes the enzymes protease, integrase, and reverse transcriptase (RT). In general, RT is an evolutionary conserved sequence (Xiong and Eickbush 1988) and the amino acid sequence is commonly used to delineate retroviral relationships. Three phylogenetic clades of LTR retrotransposons (*Gypsy/Ty3, Copia/Ty1,* and *Bel/Pao*) are currently recognized based on amino acid alignments of RT (Xiong and Eickbush 1990; Cook et al. 2000; Frame, Cutfield and Poulter 2001). Envelope (*env*) is a common retroviral feature responsible for an infectious envelope protein that surrounds the *gag*-encoded structural proteins, and is found in some LTR retrotransposons (Vogt

1997). The LTRs are identical repeats that range in size from 100-3,000 bp and flank the internal element gene regions. Regulatory signals including promoter and enhancer sequences as well as termination and polyadenylation signals are encoded by the LTRs. Since LTRs are long repeats, they may initiate a homologous recombination event resulting in the removal of the second LTR and interior sequence (Berg and Howe 1989) with a solo LTR remaining on the chromosome.

The evolution of transposon features has been predicted to be a plastic, dynamic process. For example, it is believed that the different protein domains involved with the *gag* and *pol* regions are modular (Capy et al. 1997b; Lerat et al. 1999). Phylogenetic trees of RT sequences have shown that the RT in viruses is related to the RT in transposable elements, and also related to mitochondrial intron sequences (Xiong and Eickbush 1988). Other transposon phylogenies have indicated the evolution of increasing transposon complexity, with occasional feature loss in some groups (McClure 1991; Capy et al. 1997a). Recombination may be one mechanism of domain transfer between different transposon families, in some cases leading to hybrid elements (Jordan and McDonald 1999b, a). Horizontal transfer is a process in which transposons cross into a new host species and may be important in the evolution and spread of new transposable elements features (e.g. Jordan, Matyunina and McDonald 1999; Brosius 2003). These examples point to the complex evolution of transposable elements.

Analysis of LTR retrotransposon contribution to gene function in *C. elegans* and *D. melanogaster*

Previously, studies concerning the influence of retroelements have focused most intensively on humans and mice. This dissertation extends the analysis to the influence LTR retrotransposons have upon gene evolution in the small-genome models *C. elegans* (nematode

worm) and *D. melanogaster* (fruitfly). A bioinformatics approach was used to search the sequenced genome resources of both organisms. In general, LTR retrotransposons, including fragmented elements and solo LTRs, are identified, followed by a search for the nearest neighboring genes.

Chapter 2 identifies fragmented and solo LTR TEs in the *C. elegans* genome and provides initial evidence for four TEs that may contribute to gene function (Ganko, Fielman and McDonald 2001). Several new *Cer* (*C. elegans* retrotransposon) element families and subfamilies were identified along with evidence for "bursts" of transposon activity. Many phylogenetically related families retained comparable primer binding sites (PBS). Most *Cer* elements were shown to be inserted on the chromosome arms away from the gene-rich chromosome middle.

Chapter 3 utilizes the identified *Cer* elements from Chapter 2 to identify elements in and proximal to genes in *C. elegans* (Ganko et al. 2003). A significantly greater number of transposons are found 500-1000 bp upstream of genes than predicted by random insertion models. A total of 63% of *Cer* elements are located inside genes or within the predicted regulatory boundaries of genes. Many of these genes with a nearby *Cer* element had no identified homologs, and may be novel, species-specific genes.

Chapter 4 is an analysis of LTR retrotransposons inside and proximal to genes in the sequenced euchromatin of *D. melanogaster*. One-third of identified retrotransposons are located inside genes or within the predicted regulatory boundaries of genes, and many of these retrotransposons are believed to be relatively recent insertions. Several external-reponse functional categories of genes are shown to have an overabundance of transposons, while genes with metabolic and cell differentiation functions had fewer neighboring transposons.

REFERENCES:

Ackerman, H., I. Udalova, J. Hull and D. Kwiatkowski. 2002. Evolution of a polymorphic regulatory element in interferon-gamma through transposition and mutation. Mol Biol Evol 19**:** 884-890.

Aparicio, S., J. Chapman, E. Stupka et al. 2002. Whole-Genome Shotgun Assembly and Analysis of the Genome of *Fugu rubripes*. Science**:** 1072104.

Bennetzen, J. L. 2000. Comparative sequence analysis of plant nuclear genomes:m microcolinearity and its many exceptions. Plant Cell 12**:** 1021-1029.

Berg, D. E., and M. M. Howe, 1989 *Mobile DNA*. American Society for Microbiology, Washington, D.C.

Berquin, I. M., M. Ahram and B. F. Sloane. 1997. Exon 2 of human cathepsin B derives from an Alu element. FEBS Lett 419**:** 121-123.

Boeke, J. D., D. J. Garfinkel, C. A. Styles and G. R. Fink. 1985. Ty elements transpose through an RNA intermediate. Cell 40**:** 491-500.

Boeke, J. D., and J. P. Stoye, 1997 Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements., pp. 343-436 in *Retroviruses*, edited by J. M. Coffin, S. H. Hughes and H. E. Varmus. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Bolton, E. C., and J. D. Boeke. 2003. Transcriptional interactions between yeast tRNA genes, flanking genes and Ty elements: a genomic point of view. Genome Res 13**:** 254-263.

Bowen, N. J., I. K. Jordan, J. A. Epstein, V. Wood and H. L. Levin. 2003. Retrotransposons and their recognition of pol II promoters: a comprehensive survey of the transposable elements from the complete genome sequence of *Schizosaccharomyces pombe*. Genome Res 13**:** 1984-1997.

Brosius, J. 1999. Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica 107**:** 209-238.

Brosius, J. 2003. The contribution of RNAs and retroposition to evolutionary novelties. Genetica 118**:** 99-116.

Capy, P., C. Bazin, D. Higuet and T. Langin, 1997a *Evolution and impact of transposable elements*. Kluwer Academic Publishers, Dordrecht.

Capy, P., T. Langin, D. Higuet, P. Maurer and C. Bazin. 1997b. Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? Genetica 100**:** 63-72.

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

Charlesworth, B., P. Sniegowski and W. Stephan. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371**:** 215-220.

Coffin, J. M., S. H. Hughes and H. E. Varmus, 1997 *Retroviruses*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Consortium, *C. e. S.* 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. The *C. elegans* Sequencing Consortium. Science 282**:** 2012-2018.

Conte, C., B. Dastugue and C. Vaury. 2002. Coupling of enhancer and insulator properties identified in two retrotransposons modulates their mutagenic impact on nearby genes. Mol Cell Biol 22**:** 1767-1777.

Cook, J. M., J. Martin, A. Lewin, R. E. Sinden and M. Tristem. 2000. Systematic screening of Anopheles mosquito genomes yields evidence for a major clade of *Pao*-like retrotransposons. Insect Mol Biol 9**:** 109-117.

Daborn, P. J., J. L. Yen, M. R. Bogwitz et al. 2002. A single p450 allele associated with insecticide resistance in Drosophila. Science 297**:** 2253-2256.

Dasilva, C., H. Hadji, C. Ozouf-Costaz, S. Nicaud, O. Jaillon, J. Weissenbach and H. R. Crollius. 2002. Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the *Tetraodon nigroviridis* genome. Proc Natl Acad Sci U S A 99**:** 13636-13641.

Deininger, P. L., and M. A. Batzer. 1999. Alu repeats and human disease. Mol Genet Metab 67**:** 183-193.

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

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

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

Eichler, E. E., and D. Sankoff. 2003. Structural dynamics of eukaryotic chromosome evolution. Science 301**:** 793-797.

Finnegan, D. J. 1992. Transposable elements. Curr Opin Genet Dev 2**:** 861-867.

Flavell, R. B. 1986. Repetitive DNA and chromosome evolution in plants. Philos Trans R Soc Lond B Biol Sci 312**:** 227-242.

Frame, I. G., J. F. Cutfield and R. T. Poulter. 2001. New BEL-like LTR-retrotransposons in *Fugu rubripes*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. Gene 263**:** 219-230.

Friesen, N., A. Brandes and J. S. Heslop-Harrison. 2001. Diversity, origin, and distribution of retrotransposons (*gypsy* and *copia*) in conifers. Mol Biol Evol 18**:** 1176-1188.

Gahan, L. J., F. Gould and D. G. Heckel. 2001. Identification of a gene associated with Bt resistance in *Heliothis virescens*. Science 293**:** 857-860.

Ganko, E. W., V. Bhattacharjee, P. Schliekelman and J. F. McDonald. 2003. Evidence for the Contribution of LTR Retrotransposons to *C. elegans* Gene Evolution. Mol Biol Evol 20**:** 1925-1931.

Ganko, E. W., K. T. Fielman and J. F. McDonald. 2001. Evolutionary history of *Cer* elements and their impact on the *C. elegans* genome. Genome Res 11**:** 2066-2074.

Garfinkel, D. J. 1997. Genetic loose change: how retroelements and reverse transcriptase heal broken chromosomes. Trends Microbiol 5**:** 173-175.

Goodwin, T. J., and R. T. Poulter. 2000. Multiple LTR-retrotransposon families in the asexual yeast *Candida albicans*. Genome Res 10**:** 174-191.

Green, M. M., 1988 Mobile DNA elements and spontaneous gene mutation, pp. 41-50 in *Eukaryotic transposable elements as mutagenic agents*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

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

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

Hoskins, R. A., C. D. Smith, J. W. Carlson et al. 2002. Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. Genome Biol 3**:** RESEARCH0085-0085.

Iwashita, S., N. Osada, T. Itoh et al. 2003. A transposable element-mediated gene divergence that directly produces a novel type bovine Bcnt protein including the endonuclease domain of RTE-1. Mol Biol Evol 20**:** 1556-1563.

Jordan, I. K., L. V. Matyunina and J. F. McDonald. 1999. Evidence for the recent horizontal transfer of long terminal repeat retrotransposon. Proc Natl Acad Sci U S A 96**:** 12621-12625.

Jordan, I. K., and J. F. McDonald. 1999a. Comparative genomics and evolutionary dynamics of Saccharomyces cerevisiae Ty elements. Genetica 107**:** 3-13.

Jordan, I. K., and J. F. McDonald. 1999b. Phylogenetic perspective reveals abundant Ty1/Ty2 hybrid elements in the Saccharomyces cerevisiae genome. Mol Biol Evol 16**:** 419-422.

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

Jurka, J., O. Kohany, A. Pavlicek, V. V. Kapitonov and M. V. Jurka. 2004. Duplication, coclustering, and selection of human Alu retrotransposons. Proc Natl Acad Sci U S A 101**:** 1268-1272.

Kaminker, J. S., C. M. Bergman, B. Kronmiller et al. 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol 3**:** RESEARCH0084-0084.

Kapitonov, V. V., and J. Jurka. 1999. The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. J Mol Evol 48**:** 248-251.

Kazazian, H. H., Jr., C. Wong, H. Youssoufian, A. F. Scott, D. G. Phillips and S. E. Antonarakis. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332**:** 164-166.

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

Kim, J. M., S. Vanguri, J. D. Boeke, A. Gabriel and D. F. Voytas. 1998. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res 8**:** 464-478.

Kumar, A., and J. L. Bennetzen. 1999. Plant retrotransposons. Annu Rev Genet 33**:** 479-532.

Landry, J. R., A. Rouhi, P. Medstrand and D. L. Mager. 2002. The Opitz syndrome gene Mid1 is transcribed from a human endogenous retroviral promoter. Mol Biol Evol 19**:** 1934-1942.

Le, Q. H., S. Wright, Z. Yu and T. Bureau. 2000. Transposon diversity in Arabidopsis thaliana. Proc Natl Acad Sci U S A 97**:** 7376-7381.

Leeton, P. R., and D. R. Smyth. 1993. An abundant LINE-like element amplified in the genome of Lilium speciosum. Mol Gen Genet 237**:** 97-104.

Lerat, E., F. Brunet, C. Bazin and P. Capy. 1999. Is the evolution of transposable elements modular? Genetica 107**:** 15-25.

Li, W. H., Z. Gu, H. Wang and A. Nekrutenko. 2001. Evolutionary analyses of the human genome. Nature 409**:** 847-849.

Ling, J., W. Pi, R. Bollag, S. Zeng, M. Keskintepe, H. Saliman, S. Krantz, B. Whitney and D. Tuan. 2002. The solitary long terminal repeats of ERV-9 endogenous retrovirus are conserved during primate evolution and possess enhancer activities in embryonic and hematopoietic cells. J Virol 76**:** 2410-2423.

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

Llorens, C., and I. Marin. 2001. A mammalian gene evolved from the integrase domain of an LTR retrotransposon. Mol Biol Evol 18**:** 1597-1600.

Mager, D. L., D. G. Hunter, M. Schertzer and J. D. Freeman. 1999. Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (HHLA2 and HHLA3). Genomics 59**:** 255-263.

Makalowski, W. 2000. Genomic scrap yard: how genomes utilize all that junk. Gene 259**:** 61-67.

Maside, X., C. Bartolome and B. Charlesworth. 2002. S-element insertions are associated with the evolution of the Hsp70 genes in *Drosophila melanogaster*. Curr Biol 12**:** 1686-1691.

McClintock, B. 1946. Maize genetics. Carnegie Inst. of Washington Year Book 45**:** 176-186.

McClintock, B. 1948. Mutable loci in maize. Carnegie Institute of Washington Year Book 47**:** 155-169.

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

McClintock, B. 1984. The significance of responses of the genome to challenge. Science 226**:** 792-801.

McClure, M. A. 1991. Evolution of retroposons by acquisition or deletion of retrovirus-like genes. Mol Biol Evol 8**:** 835-856.

McCollum, A. M., E. W. Ganko, P. A. Barrass, J. M. Rodriguez and J. F. McDonald. 2002. Evidence for the adaptive significance of an LTR retrotransposon sequence in a Drosophila heterochromatic gene. BMC Evol Biol 2**:** 5.

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

McDonald, J. F. 1995. Transposable elements: possible catalysts of organismic evolution. Trends Ecol. Evol 10**:** 123-126.

Medstrand, P., J. R. Landry and D. L. Mager. 2001. Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J Biol Chem 276**:** 1896-1903.

Meyers, B. C., S. V. Tingey and M. Morgante. 2001. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11**:** 1660-1676.

Morrish, T. A., N. Gilbert, J. S. Myers, B. J. Vincent, T. D. Stamato, G. E. Taccioli, M. A. Batzer and J. V. Moran. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet 31**:** 159-165.

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

Nigumann, P., K. Redik, K. Matlik and M. Speek. 2002. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics 79**:** 628-634.

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

Ostertag, E. M., and H. H. Kazazian, Jr. 2001. Biology of mammalian L1 retrotransposons. Annu Rev Genet 35**:** 501-538.

Pardue, M. L., O. N. Danilevskaya, K. Lowenhaupt, F. Slot and K. L. Traverse. 1996. Drosophila telomeres: new views on chromosome evolution. Trends Genet 12**:** 48-52.

Reiter, L. T., T. Liehr, B. Rautenstrauss, H. M. Robertson and J. R. Lupski. 1999. Localization of mariner DNA transposons in the human genome by PRINS. Genome Res 9**:** 839-843.

Richter, T. E., and P. C. Ronald. 2000. The evolution of disease resistance genes. Plant Mol Biol 42**:** 195-204.

Rostoks, N., Y. J. Park, W. Ramakrishna et al. 2002. Genomic sequencing reveals gene content, genomic organization, and recombination relationships in barley. Funct Integr Genomics 2**:** 51-59.

Satoh, N., Y. Satou, B. Davidson and M. Levine. 2003. *Ciona intestinalis*: an emerging model for whole-genome analyses. Trends Genet 19**:** 376-381.

Selker, E. U., N. A. Tountas, S. H. Cross, B. S. Margolin, J. G. Murphy, A. P. Bird and M. Freitag. 2003. The methylated component of the Neurospora crassa genome. Nature 422**:** 893-897.

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

Smit, A. F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9**:** 657-663.

Speek, M. 2001. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol 21**:** 1973-1985.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408**:** 796-815.

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

Vicient, C. M., M. J. Jaaskelainen, R. Kalendar and A. H. Schulman. 2001. Active retrotransposons are a common feature of grass genomes. Plant Physiol 125**:** 1283-1292.

Vogt, V. m., 1997 Retroviral Virions and Genomes, pp. 27-70 in *Retroviruses*, edited by J. M. Coffin, S. H. Hughes and H. E. Varmus. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Volff, J. N., C. Korting, A. Froschauer, K. Sweeney and M. Schartl. 2001. Non-LTR retrotransposons encoding a restriction enzyme-like endonuclease in vertebrates. J Mol Evol 52**:** 351-360.

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

Xiong, Y., and T. H. Eickbush. 1988. Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol Biol Evol 5**:** 675-690.

Xiong, Y., and T. H. Eickbush. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. Embo J 9**:** 3353-3362.

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

Zhou, Y. H., J. B. Zheng, X. Gu, G. F. Saunders and W. K. Yung. 2002. Novel PAX6 binding sites in the human genome and the role of repetitive elements in the evolution of gene regulation. Genome Res 12**:** 1716-1722.

Table 1.1 – Transposable element content of selected organisms

Sci. Name	Organism	genome size (Mbp)	% TE	Source
Saccharomyces cerevisiae	baker's yeast	12	3.1%	(Kim et al. 1998)
Schizosaccharomyces pombe	fission yeast	12.2	1.1%	(Bowen et al. 2003)
Candida albicans	human pathogen	16	~1.0%	(Goodwin and Poulter 2000)
Neurospora crassa	bread mold	40	~8%	(Selker et al. 2003)
Caenorhabditis elegans	C. elegans	100	6%	(Consortium 1998)
Ciona intestinalis	Sea squirt	153	~10-12%	(Satoh et al. 2003)
Drosophila melanogaster	fruitfly	165	15%	(Hoskins et al. 2002; Kaminker et al. 2002)
Anopheles gambiae	mosquito	278	>20%	(Holt et al. 2002)
Fugu rubripes	Pufferfish (marine)	365	2.70%	(Aparicio et al. 2002)
Tetraodon nigroviridis	Pufferfish (freshwater)	350	<10%	(Dasilva et al. 2002)
Mus musculus	Mouse	3,000	~34-38%	(Smit 1999; Waterston et al. 2002)
Homo sapiens	Human	3,200	> 43%	(Li et al. 2001)
Arabidopsis thaliana	Arabidopsis	130	~10%	(Le et al. 2000; The *Arabidopsis* Genome Initiative 2000)
Oryza sativa	Rice	430	<25%	(Bennetzen 2000)
Zea mays	Corn	~2,300-3,000	>55%	(Kumar and Bennetzen 1999; Meyers, Tingey and Morgante 2001)
Hordeum vulgare	Barley	4,800	>51%	(Vicient et al. 2001; Rostoks et al. 2002)
Triticum aestivum	Wheat	16,000	>90%	(Flavell 1986)
Pinus taeda	Pine	~20-30,000	>50%	(Friesen, Brandes and Heslop-Harrison 2001)
Lillium	Lilies	36,000	>95%	(Leeton and Smyth 1993)

NOTE: % TE column includes class 1 and class 2 transposons.

Table 1.2 – Examples of transposons with adaptive functionality in host genes

Transposon	Associated gene	Organism	Function	Source
LTR	Xa21 (disease resistance)	rice	regulation	(Richter and Ronald 2000)
LTR (Ty1)	Genes w/ Pol III promoters	yeast	poss. Upregulation	(Bolton and Boeke 2003)
LTR (antonia)	Cht3, chitinase 3	Drosophila	potential enhancer	(McCollum et al. 2002)
LTR (Idefix)	white locus	Drosophila	insulator	(Conte, Dastugue and Vaury 2002)
LTR (Zam)	white locus	Drosophila	enhancer	(Conte, Dastugue and Vaury 2002)
S-element	Hsp70	Drosophila	potential enhancer	(Maside, Bartolome and Charlesworth 2002)
LTR (accord)	Cyp6g1	Drosophila	5' UTR, DDT resistance	(Daborn et al. 2002)
LTR	cadherin-superfamily	Heliothis virescens (cotton pest)	Bt pesticide resistance	(Gahan, Gould and Heckel 2001)
Alu	Interferon-gamma	Human/primate	gene regulation	(Ackerman et al. 2002)
Alu	Pax6 (transcript factor)	Human	binding site of Pax6	(Zhou et al. 2002)
LINE	apolipoprotein (a)	Human	enhancer	(Yang et al. 1998)
LTR	apolipoprotein C-I	Human	alt. Promoter	(Medstrand, Landry and Mager 2001)
LTR	endothelin B receptor	Human	alt., tissue specific promotion (placental)	(Medstrand, Landry and Mager 2001)
LTR (ERV)	erythroid beta-globin locus control region (beta-LCR)	Human	alt., tissue specific promotion (embryonic and hematopoietic cells)	(Ling et al. 2002)
LTR (HERV)	leptin receptor	Human	Alt. Splicing / termination	(Kapitonov and Jurka 1999)
LTR (HERV)	HHLA3	Human	polyadenylation signal	(Mager et al. 1999)
LTR (HERV)	mid1	Human	LTR tissue specific, pot. enhancer	(Landry et al. 2002)
L1	MET- proto oncogene	Human	alt. promoter, exons	(Nigumann et al. 2002)
L1	TACTILE (T-cell surface antigen)	Human	alt. promoter, exons	(Nigumann et al. 2002)
L1	SPT3	Human	alt. promoter, exons	(Speek 2001)
Alu	cathepsin B	Human	provides exon 2	(Berquin, Ahram and Sloane 1997)
LINE (RTE-1)	Bcnt	cow	endonuclease domain	(Iwashita et al. 2003)
LTR	KIAA1051/ MyEF-3	Human / mouse	exon - domains	(Volff et al. 2001)
LTR	Gin-1	Human/mammals	gypsy-like integrase domain	(Llorens and Marin 2001)

NOTE: The last 7 examples are cases where a transposon is believed to have provided coding sequence to a host gene.

CHAPTER 2

EVOLUTIONARY HISTORY OF *CER* ELEMENTS AND THEIR IMPACT ON THE *C. ELEGANS* GENOME[1]

[1] Ganko, E., K.T. Fielman, and J.F. McDonald. 2001. *Genome Research* 11: 2066-2074. Reprinted here with permission of the publisher

ABSTRACT

We report the results of sequence analysis and chromosomal distribution of all distinguishable long terminal repeat (LTR) retrotransposons (*Cer* elements) in the *C. elegans* genome. Included in this analysis are all readily recognizable full-length and fragmented elements, as well as solo LTRs. Our results indicate that there are 19 families of *Cer* elements, some of which display significant sub-family structure. *Cer* elements can be clustered based on their tRNA primer binding sites (PBS). These clusters are in concordance with our reverse transcriptase (RT) and LTR based phylogenies. Although we find that most *Cer* elements are located in the gene depauperate chromosome ends, some elements are located in or near putative genes and may contribute to gene structure and function. The results of RT-PCR analyses are consistent with this prediction.

INTRODUCTION

Retrotransposons are an abundant and widely distributed class of mobile repetitive elements that transpose through an RNA intermediate (Berg and Howe 1989). A significant portion of eukaryotic genomes examined to date is comprised of retrotransposons. For example, more than half of the maize (>50%, (SanMiguel et al. 1996) and wheat (>90%, (Flavell 1986) genomes as well as approximately 40% of the human genome (Yoder, Walsh and Bestor 1997) is made up of retrotransposons. Long recognized as a major source of mutation (Green 1988) and disease (Miki 1998), retrotransposons have recently been implicated in the evolution of genome structure and function as well (e.g. McDonald 1995a,b; Britten 1997; Brosius 1997).

Genome sequencing of a variety of organisms is providing an unprecedented opportunity to study the evolutionary history of retrotransposons and their contribution to genome structure and function. For example, recent surveys of retrotransposons within the *C. elegans* genome have revealed the presence of no fewer than 19 families of long terminal repeat (LTR) retrotransposons (Bowen and McDonald 1999; Malik, Henikoff and Eickbush 2000; Frame, Cutfield and Poulter 2001), including two families (*Cer* 7 and *Cer* 13) which display features characteristic of infectious retroviruses (Bowen and McDonald 1999).

We extend these findings by analyzing the sequence and identifying the chromosomal distribution of all distinguishable *Cer* LTR retrotransposon sequences present in the *C. elegans* genome. In our analysis we group all distinguishable *Cer* elements into three distinct types: (1) full-length elements containing all of the characteristic features of LTR retrotransposons including putative *gag, pol* and, in some cases, *env* genes flanked by long terminal repeats (LTRs); (2) partially deleted or fragmented elements which are missing one or more of the characteristic features of full-length elements; and (3) solo LTRs which are believed to be the products of recombination events between the flanking LTRs of full-length elements (Berg and Howe 1989). Our results indicate that there are 19 *Cer* families represented within the sequenced (N2) *C. elegans* genome. All 19 families are either the *gypsy/Ty3* or *Bel* class of retrotransposons. No *copia/Ty1* type elements are present in the *C. elegans* genome.

While some full-length *Cer* elements were found to be members of extended families with well-defined evolutionary histories, others appear to be single element families with no detectable lineage within the (N2) genome. In contrast, several families of *Cer* elements were identified that are comprised of fragmented elements or solo LTRs exclusively. *Cer* elements can be grouped according to their tRNA binding sites into multiple clusters which are consistent

with our RT and LTR sequence-based phylogenies. We have also analyzed the inter- and intra-chromosomal distribution of *Cer* elements in the N2 genome. Although most *Cer* elements are located in the gene depauperate chromosomal ends, some elements are located in or near putative genes and may have contributed to gene structure and function. The results of RT-PCR analyses are consistent with this prediction. Products consistent with processing of these transcripts and removal of predicted introns were observed. *Cer* LTR sequence could account from at least 12% to as much as 54% of the coding region within mRNAs transcribed from these loci.

RESULTS:

The *C. elegans* genome consists of at least 19 families of LTR retrotransposons.

Closely related groups of full-length *Cer* LTR retrotransposons display >90% amino acid homology among their respective reverse transcriptases (RTs) and have been designated as families (Bowen and McDonald 1999). Using this criterion, full-length LTR retrotransposons representing twelve distinct families have been previously described in *C. elegans* , *Cer* 1-12 (Bowen and McDonald 1999). By searching for homology to envelop (ENV) proteins, Malik et al. (2000) discovered two additional families (*Cer* 13 and *Cer* 14). More recently, Frame et al. (2001) identified six additional putative families. In this paper, we include fragmented elements and solo LTRs in our analysis to add sub-structure to the *Cer* phylogenetic tree. Using this approach, we have independently identified a total of nineteen families (*Cer* 1-19) of *Cer* elements within the essentially complete (>99%) N2 *C. elegans* genome (*C. elegans* Sequencing Consortium 1998).

The number of *Cer* elements within families varies considerably (Table 2.1). In general, full-length *Cer* elements are in relatively low abundance within the *C. elegans* genome. Only two

of the nineteen families (*Cer* 9 and 20) contain three full-length elements while three families (*Cer* 8, 15, 16) contain two and twelve families (*Cer* 1-7,10, 12, 13, 17 and 19) contain only one. Two families (*Cer* 11,14) contain no full-length elements and another two families are comprised of only a single full-length element (*Cer* 4, 17). Fragmented elements are also in relatively low abundance (< 4 per family) in fifteen of the nineteen families (*Cer* 2, 3, 5, 6, 7, 9-16, 19, 20). Solo LTRs were detected in thirteen of the nineteen families (solo LTRs lacking in *Cer* 4, 7, 11, 13, 14, 17) ranging in number from twelve to one per family. Five of the families displayed sub-family structure. While members of *Cer* element families share >90% RT sequence identity, within family sequence identity values among the more rapidly evolving LTRs are more variable, ranging from 60-100% (Figure 2.1 and Table 2.1, 2.2).

Six of the eight solo LTRs or LTR containing fragments within the *Cer* 9 family were found to contain a ~100 bp sequence inserted into the center of their 3' LTRs (c56g3/c07d8, y57a10a and k09e3 contain a 108 bp insert; f15a2, y59a8b and c13b9 contain a 106 bp insert). Interestingly, none of the three *Cer* 9 full-length elements contain either insert within their LTRs. The ~100 bp inserts in these LTRs share 85% identity among themselves but display no significant homology to other sequences within the *C. elegans* (N2) genome. Aside from this size polymorphism, all of the *Cer* 9 family members share a remarkable 95% LTR nucleotide sequence identity with one another.

While the slowly evolving RT encoding region of LTR retrotransposons is ideal for quantitating evolutionary distances among even distantly related families of retroelements (Flavell, 1986; (Xiong and Eickbush 1990), analysis of differences among the more rapidly evolving LTRs is better suited for the identification of phylogenetic substructure within families of LTR retrotransposons. Phylogenetic trees based on *Cer* element LTR sequences reveal the presence of significant substructure within several *Cer* element families. Both neighbor-joining and parsimony

criteria support the existence of distinct subgroups in the *Cer* 2, 3, 12,15 and 16 families of elements. For example, the *Cer* 12 family is comprised of fifteen elements (primarily solo LTRs) falling into two distinct subfamilies while the fourteen elements comprising the *Cer* 16 family of elements fall into three distinct sub-families (Figure 2.2B).

Cer element families share tRNA primers.

RT requires a primer strand to initiate minus-strand DNA synthesis. Host encoded tRNA is the primer used by most retroviruses and LTR retrotransposons analyzed to date (Telsnitsky and Goff 1997). In the process of priming, the native tRNA molecule is partially unfolded such that 18 bp at its 3' terminus is free to base pair with a complementary sequence, termed the primer binding site (PBS), on the retroviral or LTR retrotransposon RNA. Different tRNA primers are known to be used by different families of retroviruses and LTR retrotransposons and have been used as an indicator of evolutionary relationships (Vogt 1997).

Primer binding sequences are located just 3' of the proviral 5' LTR. Utilizing the *C. elegans* tRNA gene database (http://rna.wustl.edu/GtRDB/Ce/), we have identified putative *Cer* primer binding sites by FASTA searches of one hundred nucleotides downstream of the 5' LTR of full-length *Cer* elements. Consistent with the recent observations of Frame et al (2001), we find that full-length elements representing *Cer* families in the *Cer* 7/BEL clade share a binding site for the Gly-GCC type (*Cer* 7-10, 12, 15, 16, 19) or for Arg type tRNAs (*Cer* 13,17, 20). We also confirm the observation of Frame et al. (2001) that *Cer* 7 encodes its own 71 bp Gly-GCC type tRNA (CE-CHRV-1298_TRNA5-GLYGCC, Figure 2.1).

Extending our alignment of putative primer binding sites and FASTA searches of the tRNA database to the *Ty3/gypsy* clade revealed that *Cer* 2 and *Cer* 3 share a PBS for Gly-ACC

tRNA. In contrast, we have found that *Cer* 4, 5 and 6 share a Ser-GCT tRNA. *Cer* 1 displays weak homology to the PBS for Thr-GGT type tRNA.

Most *Cer* elements are located at chromosome ends.

The chromosomal position of each *Cer* element was used to analyze the distribution of *Cer* elements throughout the genome. To test for inter-chromosomal clustering of *Cer* elements, we employed the Kolmogov-Smirnov goodness-of-fit test (Zarr 1999) to look for a deviation from a random distribution of elements among chromosomes. The results indicate no significant deviation from the null hypothesis (P = 0.91). The distribution of individual families of *Cer* elements (*Cer* 3, 5, 9, and 12) and family groups (Cer 8&9, P=.046; 12&16, P=.51; 2&3, P=.13) were tested separately and also found to be randomly distributed among chromosomes.

Tests were carried out to determine if the distribution of *Cer* elements on individual chromosomes was also random. Our analysis rejected the random distribution hypothesis for all chromosomes except chromosome III (Figure 2.3). Chromosomes I, II, IV, V and X were found to display non-random clustering of *Cer* elements on their chromosomal ends. This is consistent with a previous report that DNA transposable elements in *C. elegans* are clustered at chromosome ends (Surzycki and Belknap 2000) and the observation that the middle third of *C. elegans* chromosomes are "gene rich". The ends of *C. elegans* chromosomes display a lower gene density and are associated with relatively high rates of recombination (Barnes et al. 1995; Wilson 1999).

Cer elements may contribute to *C. elegans* gene function.

The results of our genomic positioning of *Cer* elements indicates that a number of these elements lie within or proximal to genes. Previous studies of LTR retrotransposons in a variety of plant and animal species have revealed that these elements may be co-opted for a variety of host gene functions, including promoter, splicing and terminator activities (Britten 1997; Medstrand, Landry and Mager 2001). In an initial effort to determine if *Cer* elements may be contributing to gene function, we screened *C. elegans* EST databases (dbEST - *C. elegans*) for homology to *Cer* elements. ESTs with significant homology to *Cer* LTRs were identified. The complete sequences of these ESTs were blasted against the *C. elegans* genome database to identify the clones containing the *Cer* LTRs and associated putative genes (F20B4, C56G3, 6R55, F53E10).

The specific region of *Cer* element identity within the four clones (F20B4.6, C5663.2, 6R55.2, F53E10.5) was overlaid on the existing annotation of each region. Our results indicate that these *Cer* elements are part of putative genes (Figure 2.4). Although all four gene regions are putative in nature, they retain strong predictive computational support. In addition, multiple ESTs were found to map to the exon regions of these putative genes adding further support. The results of TBLASTN searches indicate that two of the sites (F20B4.6 and C56G3.2) displayed significant homology (outside the *Cer* element sequence) to previously characterized genes. F20B4.6 exhibits homology to genes encoding ceramide glucosyl transferases; C56G3.2 displays homology with genes encoding aldo/keto reductases. The putative genes contained in regions 6R55.2 and F53E10.5 show no homology to genes thus far characterized (Figure 2.4).

Transcribed and Processed mRNAs contain Cer LTR sequence.

A series of reverse transcriptase polymerase chain reactions (RT-PCR) were carried out to confirm the hypothesis that *Cer* elements are contributing to the structure and function of some *C. elegans* genes. Sets of primers were designed to amplify predicted gene transcripts containing *Cer* element sequences. Because nascent RNA transcripts are typically in low abundance in standard RNA preparations, they are often underrepresented or undetectable in the products of RT-PCR reactions. For this reason, PCR of genomic DNA was also carried out for each set of primers as a positive control.

Primers designed for the 6R55.2 gene yielded RT-PCR products consistent with the expected sizes of the nascent (1514 bp) and processed (429 bp) transcripts (Figure 2.5). A 6R55.2 transcript fully processed according to its predicted gene structure (Figure 2.4A) would contain 16% LTR sequence from a *Cer* 16-2 element in its coding region. If all exons represented by EST alignments (Figure 2.4A) were present in the final processed transcript, 54% of its coding region would be LTR sequence.

Primers designed for the C56G3.2 gene yielded RT-PCR products consistent with the expected size of the nascent (634 bp) and processed (569 bp) transcripts (Figure 2.5). The smaller RT-PCR product is consistent with excision of the intron predicted within the *Cer* 9 LTR. It is intriguing to note that the position of the predicted intron within the Cer 9 LTR overlaps with an approximate 100 bp sequence missing in some of the solo LTRs identified in this study. A 6R55.2 transcript fully processed according to its predicted gene structure (Figure 2.4B) would have first and second exons comprised of 100% and 40% of *Cer* 9 LTR sequence, respectively. Thus, within its coding region the mRNA would be 36% LTR sequence. Alternate

intron/exon structures (Figure 2.4B) could generate transcripts ranging from 20 to 48% *Cer* 9 LTR as mRNA coding sequence.

Primers designed for the F20B4.6 gene yielded a preferentially amplified RT-PCR product of ~213 bp. This product is consistent with excision of the *Cer* 16-1 LTR from intron 1 (Figs. 2.4C and 2.5), although potential enhancer activity of the LTR cannot be excluded by this analysis. Two bands at ~380 and ~430 bp may represent unpredicted processing products or non-specific priming, although they were also apparent in reactions performed at temperatures 10°C higher than the predicted optimum for the pair (data not shown).

Primers designed for the F53E10.5 yielded two RT-PCR products consistent with predicted processing of the nascent transcript (Figure 2.5). A weakly amplified product at ~520 bp is consistent with mRNA processing and removal of intron 9 (Figure 2.4D). The preferentially amplified product at ~449 bp is consistent with removal of introns 8 and 9. Exon 10, derived entirely from *Cer* 2 LTR DNA, would contribute 12% coding sequence if the mRNA was fully processed as predicted.

In summary, RT-PCR analyses demonstrated that the inserted *Cer* elements were part of each gene transcript, thus providing molecular confirmation of our computational results (Figure 2.5). Polyadenylated transcripts composed of retroelement sequence were produced from the three genes in which elements were part of the coding region. Furthermore, products consistent with processing of these transcripts and removal of predicted introns were observed.

DISCUSSION

The *C. elegans* genome contains relatively few families of LTR retrotransposons with unusual sub-family structure.

Nucleotide sequence divergence among LTR retrotransposons can be used to establish phylogenetic relationships and other relevant information related to retrotransposon evolution. Our approach has been to utilize RT sequence to establish families (defined as groups of LTR-retrotransposons sharing at least 90% RT sequence homology) and to subsequently utilize the divergence among the more rapidly evolving LTRs to establish sub-family structure. An alternative approach recently employed by Frame et al (2001) to characterize the BEL-like class of *C. elegans* LTR retrotransposons, is to base phylogenetic relationships primarily upon LTR sequences. *A priori*, both approaches might be expected to give similar results. However, because the *C. elegans* genome contains relatively few full-length elements and relatively more fragmented elements and solo LTRs lacking RT sequences, the former approach will tend to identify fewer families of elements with more sub-structure than the latter approach. For example, the *Cer* 16 *and Cer* 18 families of Frame et al.(2001) are collapsed in our analysis to a single family (*Cer* 16) with detailed sub-family structure. As more data become available on the diversity of LTR retrotransposons present in other strains of *C. elegans,* the results should converge on a single picture of the evolutionary history of *Cer* elements.

Although our view of the phylogenetic structure of *Cer* elements differs somewhat from that recently described by Frame et al. (2001), we find that many of the general features of the *Cer 7* / BEL class of *C. elegans* LTR retrotransposons described by these authors hold true for the *Ty3/gypsy* class as well. In general, the *C. elegans* genome appears to have a relatively low tolerance for LTR retrotransposons (<1%). While we have identified 124 full-length, fragmented

or solo LTR *Cer* elements in the sequenced (N2) *C. elegans* genomes, >350 LTR retrotransposon elements have been described in the yeast *Candida albicans* (Goodwin and Poulter 2000) and >300 in *Saccharomyces cerevisiae* (Kim et al. 1998), both species with genomes nearly an order of magnitude smaller than *C. elegans* (Consortium 1998).

Single element groups add to the puzzle. Families represented by only one element (*Cer* 4, 11, 17) have no detectable history in the *C. elegans* (N2) genome, suggesting that they may have been introduced by horizontal transfer. The fact that the *Cer* 7 and *Cer* 14 elements encode a putative *env* gene is consistent with the hypothesis that at least some *Cer* elements may have entered the N2 genome via horizontal transfer. However, additional information on the diversity of elements in other *C. elegans* strains and related *Caenorhabditis* species will be necessary in order to definitively test the horizontal transfer hypothesis.

A number of solo LTRs and LTR containing fragments are nearly identical in sequence despite the fact that related full-length putative progenitor elements are not present in the genome. For example, the *Cer* 3-1 sub-family consists of ten solo LTRs and one LTR-containing fragment with > 94% identity. Similarly, the *Cer* 16-1 sub-family consists of 6 solo LTRs with >94% identity. Despite the sequence similarity among these and other sub-family LTRs, the sequences of *Cer* 16-1 LTRs are distinctly different from their most closely related full-length elements. One possible explanation of this apparent paradox is that some mechanism exists in *C. elegans* to rapidly remove full-length transposable elements as has been postulated in *Drosophila* (Petrov, Lozovskaya and Hartl 1996). Under this scenario, solo LTRs and LTR-containing fragments are remnants of degraded full-length elements. Alternatively, the high sequence similarity existing among families of solo LTRs and LTR-containing fragments may be the product of gene conversion.

A third possible explanation is that at least some of the families of solo LTRs and LTR-containing fragments represent footprints of double-strand break (DSB) repair events (Garfinkel 1997). Teng et al. (1996) and Yu and Gabriel (1999) have reported that a variety of *Ty1* LTR transcription intermediates have been used to repair double stranded breaks in *Saccharomyces cerevisiae*. If such a mechanism exists in *C. elegans*, it is possible that at least some sub-families of LTRs displaying high sequence similarity may have been copied off of the same master element during the process of DSB repair.

The presence of tRNA genes in *Cer* elements may be of adaptive significance.

Putative tRNA primer-binding sites have been identified for most full-length *Cer* elements. Matching tRNAs consisted predominantly of glycine (TCC and ACC) types. The distribution of these different types of tRNA binding sites was found to be consistent with our RT based phylogeny (Figure 2.1).

It is interesting to speculate on the significance of the surprising finding that a complete tRNA-Gly gene is located within the untranslated leader region of *Cer 7*. The observation that LTR retrotransposons are common in heterochromatic regions of genomes (Dimitri and Junakovic 1999) has led to the speculation that the evolutionary origin of heterochromatin may have been as a defense mechanism against transposable elements (McDonald 1999; Henikoff 2000). tRNA genes are known to exclude nucleosomes and limit the spread of heterochromatin (Morse 2000). Thus, the inclusion of a tRNA gene in an LTR retrotransposon may provide a selective advantage to an element located in heterochromatic regions by preventing nucleosome positioning. The consequent exclusion of surrounding chromatin may permit access of transcription factors to promoter sequences within the LTR and adjacent leader regions that

would otherwise be inaccessible. Although the *C. elegans* genome does not contain consitutive heterochromatin, transient heterochromatin-like structures occur during development (e.g. Jedrusik and Schulze 2001). As analyses of LTR retrotransposons are extended to additional plant and animal species, it will be interesting to see if the presence of complete tRNA genes in untranslated leader regions is a general feature of some families of LTR retrotransposons.

Cer elements may contribute to *C. elegans* gene structure and function.

There is a growing body of evidence that transposable elements may play an important role in genome evolution by contributing to the structure and/or function of genes (e.g., McDonald 1995a,b; (Britten 1997; Medstrand, Landry and Mager 2001). For example, there are over one hundred reported examples of essential gene structures and functions in mammals that are attributable to retrotransposons or retrotransposon derived sequences (Brosius 1999). LTRs are known to possess promoter, polyadenylation and enhancer functions (e.g., Medstrand, et al. 2001; Britten 1997). For this reason, LTR retrotransposon insertions in or near genes have been postulated to be a significant factor in regulatory evolution in both plants and animals (e.g., (McDonald 1993; McDonald 1995). The insertion of transposable elements in or near introns can result in alternative splicing patterns. Such events are also believed to have contributed to gene evolution (Kapitonov and Jurka 1999). The insertion of transposable elements into the coding region of genes is typically associated with loss of gene function (Green 1988). However, occasionally such events are associated with alterations in gene sequence which may contribute to the evolution of new gene functions (e.g., (Banki, Halladay and Perl 1994).

In an initial effort to address the possible contribution of *Cer* elements to *C. elegans* gene evolution, we screened *C. elegans* EST databases for the presence of *Cer* element LTRs.

We have identified four genes in which *Cer* elements may be involved in gene function. In three cases, LTR sequences appear to be incorporated into coding regions (Figure 2.4A,B,D). In addition, we have found that *Cer* LTRs map to putative gene splice acceptor/donor sequences and termination regions of genes (Figure 2.4A,B,C). These results are intriguing and suggest that *Cer* LTRs may influence gene regulation and expression in the *C. elegans* N2 strain.

RT-PCR analyses confirmed that mRNAs containing *Cer* LTR sequence are actively transcribed from these loci. In three of the four loci, *Cer* element sequences mapped to coding regions of the genes. For each of these cases, polyadenylated transcripts were shown to be produced containing the expected *Cer* LTR (Figure 2.5). Furthermore, products consistent with processing of these transcripts and removal of predicted introns were also observed. *Cer* LTR sequences could account from at least 12% to as much as 54% of the coding region within mRNAs transcribed from these loci. Detailed molecular analyses are currently underway in our laboratory to precisely define the contribution of *Cer* elements to the function of these genes in the N2 strain and to examine the functional significance of *Cer* element insertional polymorphisms at these and other loci among *C. elegans* strains.

METHODS:

Sequence Identification and Retrieval

Sequence retrieval was initiated by performing BLASTN searches (default parameters-(Altschul et al. 1997)) against the Wormbase (www.wormbase.org) and GenBank (www.ncbi.nlm.nih.gov) databases using LTRs representing each previously identified family of *Cer* elements (Bowen and McDonald 1999; Malik, Henikoff and Eickbush 2000). In order to

insure that all families of *Cer* LTRs were identified, we employed an iterative approach whereby LTR sequences with relatively low homology (~70%) were used as query sequences in subsequent BLAST searches to identify putative distantly related sub-families of LTRs. To be considered an LTR in this study, a sequence had to display >60% sequence homology to the LTR query sequence in a pairwise comparison test (Tatusova and Madden 1999) and have a size no smaller than 40% of the LTR query sequence. Each *Cer* LTR identified by these criteria was given the name of the *Cer* family to which it was most homologous followed by the number of the clone in which it was found. For full-length elements having two LTRs, the 3' LTR is labeled by a lower case "b" following the clone number.

Alignments and phylogenetic analysis

Using the clone coordinates from the BLAST search, the *Cer* LTR sequences were copied and placed into individual files. Alignments were created with ClustalW and edited with MacVector 7.0 (www.gcg.com). Both Clustalx 1.8 (Thompson et al. 1997) and PAUP 4.03b (Swofford 1999) were used to generate NJ trees with bootstrap values. Trees were viewed with TreeView 1.5.3 (Page 1996).

tRNA identification

The *C. elegans* tRNA database was downloaded (http://rna.wustl.edu/tRNAdb/; Lowe and Eddy 1997) for use as a local FASTA database in conjunction with the GCG software package (www.gcg.com) maintained by the Research Computing Resource (RCR) at the University of Georgia. One hundred and one nucleotides downstream of each 5' LTR (including

the last nucleotide of the LTR) were used as query sequences in FASTA searches (default parameters) run against the tRNA database to identify matching tRNA 3' ends complementary to putative *Cer* PBS (Goodwin and Poulter 2000).

Chromosomal position analyses

The chromosomal position of the 5' end of each clone found to contain one or more *Cer* elements was obtained from Wormbase (www.wormbase.org). Endpoints of elements within clones were averaged to obtain a "position value" for each element within a clone. Combining position values of elements within a clone with the position of clones on chromosomes allowed us to assign a chromosomal location to each *Cer* element. The Kolmogov-Smirnov goodness-of-fit test was used to test the randomness of the distribution of *Cer* elements among chromosomes and within individual chromosomes. An exponential distribution was used to represent a random dispersal of elements within each chromosome. The observed distribution was calculated based on the base pair distance between sequential element positions along the chromosome.

Gene annotation

The *C. elegans* EST database (dbEST - *C. elegans*) was BLASTed for homology to each *Cer* LTR sequence. ESTs with significant homology (e <0.0001) to *Cer* LTRs were identified. The complete sequences of each EST were then BLASTed against the NCBI *C. elegans* genome database to identify the corresponding clone containing the LTR and associated gene. TBLASTN searches (default parameters) of these LTR associated genes were run to identify

homology to previously characterized genes. GeneFinder (dot.imgen.bcm.tmc.edu) was used to delineate the exon boundaries of the putative genes.

RT-PCR

Total RNA was extracted with Tri Reagent® (Molecular Research Center) from *C. elegans* cultured under standard conditions on mixed life stage agar plates (Wood 1988). DNA contamination was removed using DNA-*free* (Ambion). Oligo dT_{20} primed reverse transcription (RT) was performed on 1 μg of total RNA using the ThermoScript RT-PCR system and protocol from Gibco BRL. RT (-) control reactions to detect DNA contamination contained an equivalent volume of sterile distilled water in lieu of reverse transcriptase.

PCR primers designed with MacVector 7.0 and synthesized by Integrated DNA Technologies were: 6R55.2 F 5'-ATGACGATGAGCGGTGC-3', R 5'-AAAGTGAGATGTGATTGGGG-3'; C56G3.2 F 5'-CAGCAACCTTCCTACACGG-3', R 5'-CGCAACTCAGATGGAGCAG-3'; F20B4.6 F 5'-AAGGGTTGGGTTTGGTTGGAC-3', R 5'-TCAAGAACAGAACGCCTCGTCG-3'; and F53E10.5 F 5'-GCGATAGCGTTCTGCTCTTGTG-3', R 5'-GGCGAATAAATGAAATCACGGAGG-3' (Figure 2.4). Within a locus, PCRs on genomic DNA and cDNAs were performed using the same primer set. The 25 μl PCRs contained 2 μl RT reaction or *C. elegans* genomic DNA, 30 pmol of each primer, 0.5 U *Taq* polymerase (Pierce Chemical Company), 200 µM each dNTP, 1.5 mM $MgCl_2$, 50 mM KCl, and 10 mM Tris HCl, pH 9.0. DNA (-) PCR controls to detect potential DNA contamination contained an equivalent volume of sterile distilled water in lieu of genomic DNA. Following an initial denaturation at 95 ºC / 5 min, 35 cycles of 95 ºC / 30 sec, 52 to 56 ºC (primer dependent) / 30 sec, 72 ºC / 1 to 2 min (depending on maximum expected

product length), and a final cycle at 72 ºC for 10 minutes were performed on a Hot Top equipped RoboCycler® Gradient 96 (Stratagene). Reaction products (15 μl) and a 100 bp ladder (0.25 μg)(New England Biolabs) were separated on a 1.3% agarose gel in 0.5 x TBE running buffer containing 0.25 μg ml^{-1} ethidium bromide. Gel images were visualized by UV transillumination and scanned for image processing.

ACKNOWLEDGMENTS

Thanks to King Jordan for statistical advice and Nathan Bowen for reading and commenting on earlier versions of this manuscript. Our laboratory is supported by grants from the National Institutes of Health and the National Science Foundation.

REFERENCES

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

Banki, K., D. Halladay and A. Perl. 1994. Cloning and expression of the human gene for transaldolase. A novel highly repetitive element constitutes an integral part of the coding sequence. J Biol Chem 269**:** 2847-2851.

Barnes, T. M., Y. Kohara, A. Coulson and S. Hekimi. 1995. Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. Genetics 141**:** 159-179.

Berg, D. E., and M. M. Howe, 1989 *Mobile DNA*. American Society for Microbiology, Washington, D.C.

Bowen, N. J., and J. F. McDonald. 1999. Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. Genome Res 9**:** 924-935.

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

Brosius, J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238**:** 115-134.

Consortium, *C. e. S.* 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282**:** 2012-2018.

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

Flavell, R. B. 1986. Repetitive DNA and chromosome evolution in plants. Philos Trans R Soc Lond B Biol Sci 312**:** 227-242.

Frame, I. G., J. F. Cutfield and R. T. Poulter. 2001. New BEL-like LTR-retrotransposons in *Fugu rubripes*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. Gene 263**:** 219-230.

Garfinkel, D. J. 1997. Genetic loose change: how retroelements and reverse transcriptase heal broken chromosomes. Trends Microbiol 5**:** 173-175.

Goodwin, T. J., and R. T. Poulter. 2000. Multiple LTR-retrotransposon families in the asexual yeast *Candida albicans*. Genome Res 10**:** 174-191.

Green, M. M., 1988 Mobile DNA elements and spontaneous gene mutation, pp. 41-50 in *Eukaryotic transposable elements as mutagenic agents*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Henikoff, S. 2000. Heterochromatin function in complex genomes. Biochim Biophys Acta 1470**:** 1-8.

Kapitonov, V. V., and J. Jurka. 1999. The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. J Mol Evol 48**:** 248-251.

Kim, J. M., S. Vanguri, J. D. Boeke, A. Gabriel and D. F. Voytas. 1998. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. Genome Res 8**:** 464-478.

Malik, H. S., S. Henikoff and T. H. Eickbush. 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10**:** 1307-1318.

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

McDonald, J. F. 1995. Transpoasble elements: possible catalysts of organismic evolution. Trends Ecol. Evol 10**:** 123-126.

McDonald, J. F. 1999. Genomic imprinting as a co-opted evolutinary character. Trends Ecol. Evol 13**:** 94-95.

Medstrand, P., J. R. Landry and D. L. Mager. 2001. Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J Biol Chem 276**:** 1896-1903.

Miki, Y. 1998. Retrotransposal integration of mobile genetic elements in human diseases. J Hum Genet 43**:** 77-84.

Morse, R. H. 2000. RAP, RAP, open up! New wrinkles for RAP1 in yeast. Trends Genet 16**:** 51-53.

Page, R. D. 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12**:** 357-358.

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

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

Surzycki, S. A., and W. R. Belknap. 2000. Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. Proc Natl Acad Sci U S A 97**:** 245-249.

Swofford, D. L., 1999 PAUP* Phylogenetic analysis using parsimony (* and other methods), pp. Sinuaer Assoc., Sunderland, MA.

Tatusova, T. A., and T. L. Madden. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174**:** 247-250.

Telsnitsky, A., and S. P. Goff, 1997 Reverse Transcriptase and the Generation of Retroviral DNA, pp. 121-161 in *Retroviruses*, edited by J. M. Coffin, S. H. Hughes and H. E. Varmus. Cold Spring Harbor Laboratory Press, New York.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25**:** 4876-4882.

Vogt, V. M., 1997 Reverse Transcriptase and the Generation of Retroviral DNA, pp. 27-70 in *Retroviruses*, edited by J. M. Coffin, S. H. Hughes and H. E. Varmus. Cold Spring Harbor Laboratory Press, New York.

Wilson, R. K. 1999. How the worm was won. The *C. elegans* genome sequencing project. Trends Genet 15**:** 51-58.

Xiong, Y., and T. H. Eickbush. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. Embo J 9**:** 3353-3362.

Yoder, J. A., C. P. Walsh and T. H. Bestor. 1997. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13**:** 335-340.

Zarr, J., 1999 *Biostatistical analysis*. Prentice Hall, Upper Saddle River, N.J.

Table 2.1: Number of full length, fragmented and solo LTRs in the sequenced *C. elegans* (N2) genome.

Cer Element	Full Length	Fragment	Solo LTR	Subtotals
Cer1	1	0	3	4
Cer2	1	0	1	2
Cer2-1	0	2	1	3
Cer3	1	0	0	1
Cer3-1	0	1	10	11
Cer4	1	0	0	1
Cer5	1	4	9	14
Cer6	1	1	4	6
Cer7	1	1	0	2
Cer8	2	0	1	3
Cer9	3	3	5	11
Cer10	1	1	3	5
Cer11	0	1	0	1
Cer12	1	0	10	11
Cer12-1	0	2	2	4
Cer 13	1	1	0	2
Cer 14	0	1	0	1
Cer15	1	1	1	3
Cer15-1	1	2	1	4
Cer16	1	0	3	4
Cer16-1	0	0	6	6
Cer16-2	1	3	1	5
Cer 17	1	0	0	1
Cer 19	1	4	6	11
Cer 20	3	1	4	8
Totals	24	29	71	124

Table 2.2: List of all known *Cer* LTR retrotransposons in the sequenced *C. elegans* (N2) genome.

Genomic clone ID and chromosome locations were obtained from Wormbase (www.wormbase.org).

Element Family	Genomic clone	Element Type	Chromosome
Cer1	f44e2/par3	full	III
Cer1	c25a11	ltr	X
Cer1	c24h10	ltr	X
Cer1	y39e4b	ltr	III
Cer2	r03d7	full	II
Cer2	f53e10	ltr	V
Cer2-1	k08d10	frag	IV
Cer2-1	f49f1	frag	IV
Cer2-1	w04a8	ltr	I
Cer3	f58h7	full	IV
Cer3-1	y37h2a	frag	V
Cer3-1	y76b12c	ltr	IV
Cer3-1	y39e4a	ltr	III
Cer3-1	k09h9	ltr	I
Cer3-1	y39b6a	ltr	V
Cer3-1	e02h9	ltr	III
Cer3-1	y105e8a	ltr	I
Cer3-1	y23h5b	ltr	I
Cer3-1	y77e11a	ltr	IV
Cer3-1	y75b8a	ltr	III
Cer3-1	t09a5	ltr	II
Cer4	f15g10/t23e7	full	X
Cer5	t03f1	full	I
Cer5	f39b3	frag	X
Cer5	k02a2	frag	II
Cer5	c31e10	frag	X
Cer5	f22g12	frag	I
Cer5	r01h5	ltr	X
Cer5	y27f2a	ltr	II
Cer5	t27f6	ltr	I
Cer5	c25b8	ltr	X
Cer5	f49c8	ltr	IV
Cer5	f22e5	ltr	II
Cer5	w04g5	ltr	I
Cer5	y111b2a	ltr	III

Cer5	f56h6	ltr	I
Cer6	e03a3	full	III
Cer6	y102a5c	frag	V
Cer6	y53f4a	ltr	II
Cer6	y73f8a	ltr	IV
Cer6	zc487	ltr	V
Cer6	c55a1	ltr	V
Cer7	zc132	full	V
Cer7	h08m01	ltr	IV
Cer8	zk262/zk228	full	V
Cer8	c03a7	full	V
Cer8	c33e10	ltr	X
Cer9	y43f4a	full	III
Cer9	w09b7 / f07b7	full	V
Cer9	f07b7 / k06c4	full	V
Cer9	k09e3	frag	X
Cer9	c33c12/c40a11	frag	II
Cer9	c07d8 / c56g3	frag	X
Cer9	b0047	ltr	II
Cer9	c13b9	ltr	III
Cer9	y59a8b	ltr	V
Cer9	y57a10a	ltr	II
Cer9	f15a2	ltr	X
Cer10	y81b9a/c35b8	full	X
Cer10	t23b12/zk994	frag	V
Cer10	t12b5	ltr	III
Cer10	y73f8a	ltr	IV
Cer10	t22b2	ltr	X
Cer11	t14g12	frag	X
Cer12	f21d9/f55c9	full	V
Cer12	k07c6	ltr	V
Cer12	w03g1	ltr	IV
Cer12	y51h4a	ltr	IV
Cer12	c01b9	ltr	II
Cer12	c09g1	ltr	X
Cer12	k04c1	ltr	X
Cer12	c44b12	ltr	IV
Cer12	y94h6a	ltr	IV
Cer12	c04g6	ltr	II
Cer12	y60a3a	ltr	V
Cer12-1	zc15	frag	V
Cer12-1	f41g4	frag	X
Cer12-1	k08d12	ltr	IV
Cer12-1	f58f6	ltr	IV
Cer13	y75d11a/w03h1	full	X
Cer13	c09b9	frag	IV
Cer14	y105c5b	frag	IV
Cer15	y102a5d/f40d4	full	V
Cer15	t11f9	frag	V

Cer15	y105e8a	ltr	I
Cer15-1	c52e2/ c16c4	full	II
Cer15-1	f19b2	frag	V
Cer15-1	y40h7a	frag	IV
Cer15-1	y45f10c	ltr	IV
Cer16	r13d11	full	V
Cer16	f47d2	ltr	V
Cer16	f28d9	ltr	I
Cer16	f36a4	ltr	IV
Cer16-1	y71h2am	ltr	III
Cer16-1	f38c2	ltr	IV
Cer16-1	f11a6	ltr	I
Cer16-1	y32h12a	ltr	III
Cer16-1	f47b7	ltr	X
Cer16-1	f20b4	ltr	X
Cer16-2	f20b4 / 6r55	full	X
Cer16-2	zk1025 / c27c7	frag	I
Cer16-2	t16g12	frag	III
Cer16-2	t05a1	frag	IV
Cer16-2	zc247	ltr	I
Cer17	r52	full	II
Cer19	r09h3 / c36c9	full	X
Cer19	c38d9	frag	V
Cer19	y7a5a	ltr	X
Cer19	f15d4	frag	II
Cer19	t06a10	frag	IV
Cer19	d1022	ltr	II
Cer19	f35h10	ltr	IV
Cer19	zk1055	ltr	V
Cer19	c35d6	ltr	IV
Cer19	t08g3	ltr	V
Cer19	zk643	ltr	III
Cer20	y87g2a	full	I
Cer20	k01d12	full	V
Cer20	f41b5	frag	V
Cer20	y50e8a	frag	V
Cer20	t10g3	ltr	V
Cer20	t28d6	ltr	III
Cer20	c32b5	ltr	II
Cer20	y94h6a	ltr	IV

Figure 2.1: Composite RT / LTR phylogenetic analysis of *Cer* elements.

Shown is an unrooted NJ phylogram of RT (amino acid) and LTR (nucleotide) sequences. RT amino acid alignments were used to establish family structure (black); LTR nucleotide sequences were added to establish subfamily structure (red). tRNA primer binding sites (PBS) are highlighted to show conservation of tRNA priming across families. Alignments were produced via MacVector (www.gcg.com) and Clustalx 1.8 (Thompson et al. 1997). * PBS reported by Frame, et al. (2001).

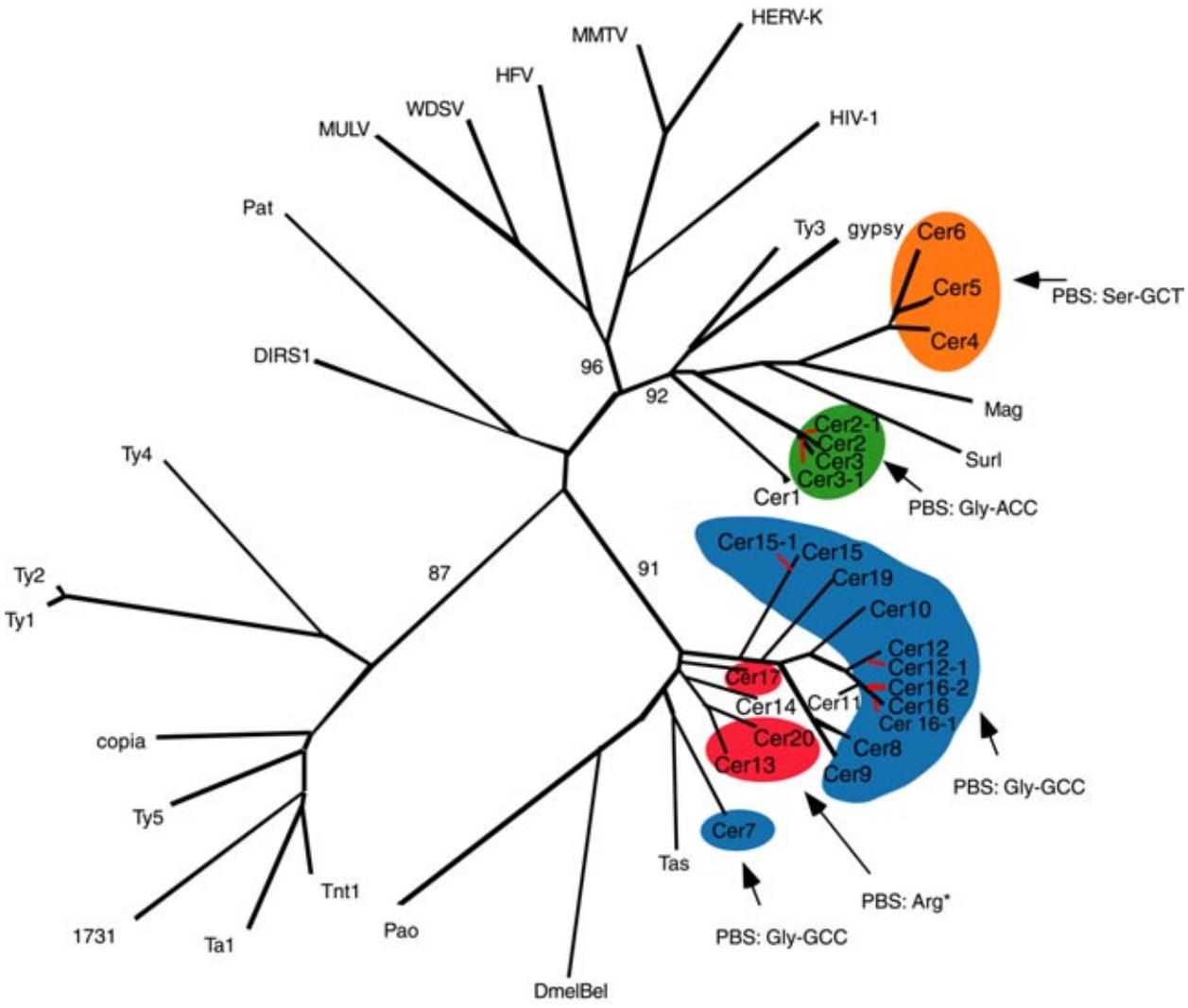

HERV-K
MMTV
HFV
WDSV
MULV
HIV-1
Pat
Ty3
gypsy
Cer6
Cer5
Cer4
PBS: Ser-GCT
DIRS1
96
92
Mag
Surl
Cer2-1
Cer2
Cer3
Cer3-1
Cer1
PBS: Gly-ACC
Ty4
Cer15-1
Cer15
Cer19
Cer10
Ty2
Ty1
87
91
Cer12
Cer12-1
Cer16-2
Cer16
Cer 16-1
Cer17
Cer14
Cer11
Cer20
Cer13
Cer8
Cer9
PBS: Gly-GCC
copia
Ty5
Cer7
Tas
Tnt1
PBS: Arg*
1731
Ta1
Pao
PBS: Gly-GCC
DmelBel

Figure 2.2: Phylogenetic trees of sub-family structure based on LTR nucleotide sequence data. LTRs from full, fragmented, and solo LTR elements were aligned via Clustalx 1.8 (Thompson et al. 1997) and the NJ method was used to construct trees. Insertions/deletions were ignored. Values on individual branches are bootstrap percentages based on 1000 bootstrap repetitions. Each LTR in the tree is named by the genomic clone in which it was found. For elements with two LTRs the 3' LTR is labeled by a lower case "b" following the clone number. Each tree is shown with a scale bar determined by the number of nucleotide substitutions per site between two sequences.

A: Phylogenetic tree displaying sub-structure within *Cer* 8 & 9 families, with *Cer* 7 as the outgroup. The tight branching of the tree demonstrates the high sequence identity shared among *Cer* 9 family members. * indicates the presence of a ~108 bp insert in the center of *Cer* 9 LTR; ** indicates the presence of a ~106 bp insert in the center of the *Cer* 9 LTR. Both inserts are >85% identical.

B: Phylogenetic tree displaying sub-structure within *Cer* 12 & 16 families, with *Cer* 7 as the outgroup. *Cer* 12 consists of 2 subfamilies (*Cer* 12 & 12-1), *Cer* 16 has 3 subfamilies (*Cer* 16, 16-1, 16-2). The tight clustering seen in both families represents a high degree of nucleotide identity between elements within a subfamily.


A
cer9 c56g3 *
100
cer9 c07d8 *
cer9 b0047
cer9 k09e3b *
cer9 y57a10a *
cer9 y43f4a
cer9 k06c4
97
cer9 w09b7
90
cer9 k09e3
cer9 c40a11
98
100
cer9 c33c12
cer9 f15a2 **
cer9 y59a8b **
100
cer9 c13b9 **
60
cer8 c33e10
78
cer8 zk262
100
46
cer8 c03a7b
100
cer8 c03a7
cer7 zc132
0.1

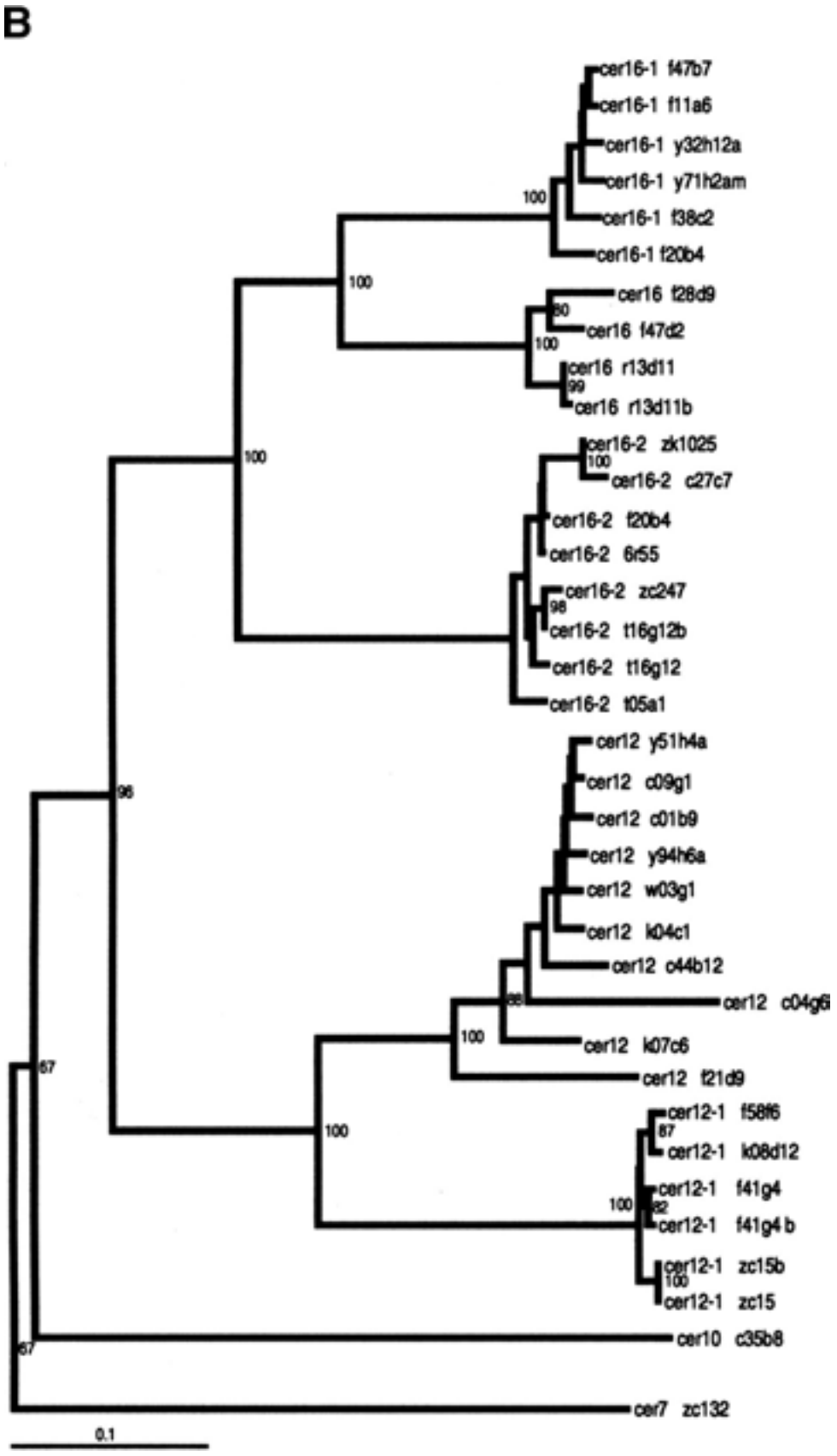

B
cer16-1 f47b7
cer16-1 f11a6
cer16-1 y32h12a
cer16-1 y71h2am
cer16-1 f38c2
cer16-1 f20b4
cer16 f28d9
cer16 f47d2
cer16 r13d11
cer16 r13d11b
cer16-2 zk1025
cer16-2 c27c7
cer16-2 f20b4
cer16-2 6r55
cer16-2 zc247
cer16-2 t16g12b
cer16-2 t16g12
cer16-2 t05a1
cer12 y51h4a
cer12 c09g1
cer12 c01b9
cer12 y94h6a
cer12 w03g1
cer12 k04c1
cer12 c44b12
cer12 c04g6
cer12 k07c6
cer12 f21d9
cer12-1 f58f6
cer12-1 k08d12
cer12-1 f41g4
cer12-1 f41g4 b
cer12-1 zc15b
cer12-1 zc15
cer10 c35b8
cer7 zc132
0.1

Figure 2.3: Distribution of *Cer* full-length, fragmented and solo LTR element sequences in the *C. elegans* genome.

A genomic coordinate value for all *Cer* elements was calculated (see Methods) and elements plotted to their respective chromosome location. Chromosomes were divided into three regions (left, centric, right). All chromosomes except chromosome III display statistically significant clustering outside of the centric genic region. *Cer* elements are randomly distributed across chromosomes.

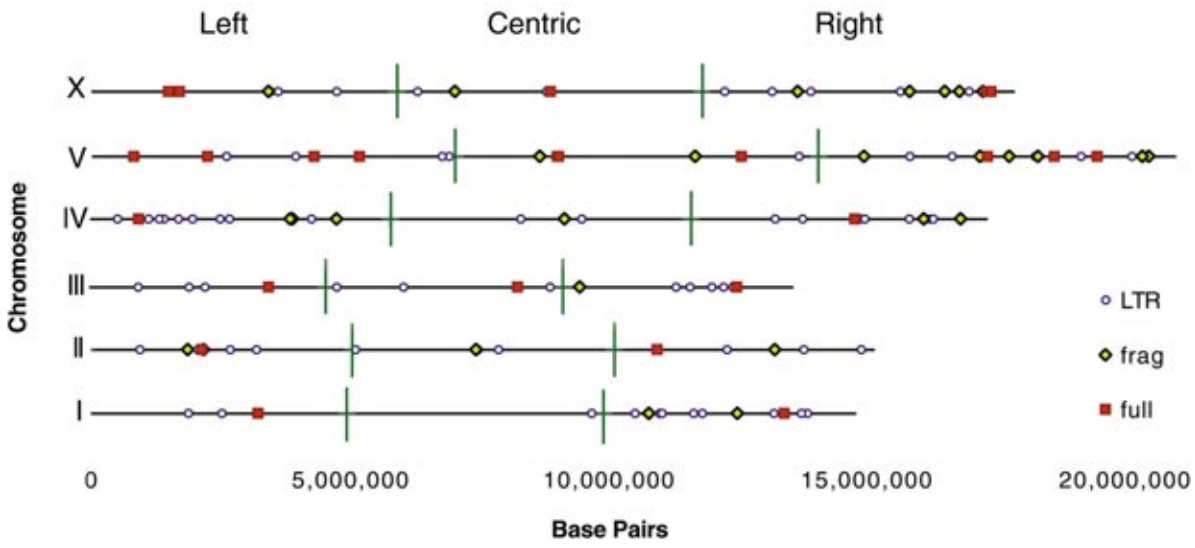

Left
Centric
Right
X
V
IV
III
II
I
Chromosome
LTR
frag
full
0
5,000,000
10,000,000
15,000,000
20,000,000
Base Pairs

Figure 2.4: *Cer* element LTRs are part of some *C. elegans* genes.

Green arrows represent Wormbase-predicted gene regions with corresponding identification. Blue arrows depict ESTs concordant to the predicted gene region. Orange boxes are predicted exon regions. Red boxes denote LTR position and internal arrows indicate direction. The black line and numbers represent position along the genomic clone sequence (F20B4, C56G3, 6R55, F53E10). Black arrows indicate direction and location of forward (f) or reverse(r) PCR primers. For visual simplicity, only introns (i#) discussed in the text are displayed above and between exons.

A. An entire LTR from the 5' end of a full-length 16-2 element is part of the 5' end of a putative *C. elegans* gene (6R55.2) of unknown function.

B. The *Cer* 9 LTR overlaps 2 exons of an aldo/keto reductase homolog in *C. elegans* (C56G3.2). The LTR is the 3' end of a fragmented *Cer* 9 element.

C. A *Cer*16-1 solo LTR is part of intron 1 of a *C. elegans* gene (F20B4.6) in the glucosyltransferase family.

D. A *Cer* 2 solo LTR constitutes the 3' end of a putative *C. elegans* gene (F53E10.5) of unknown function.

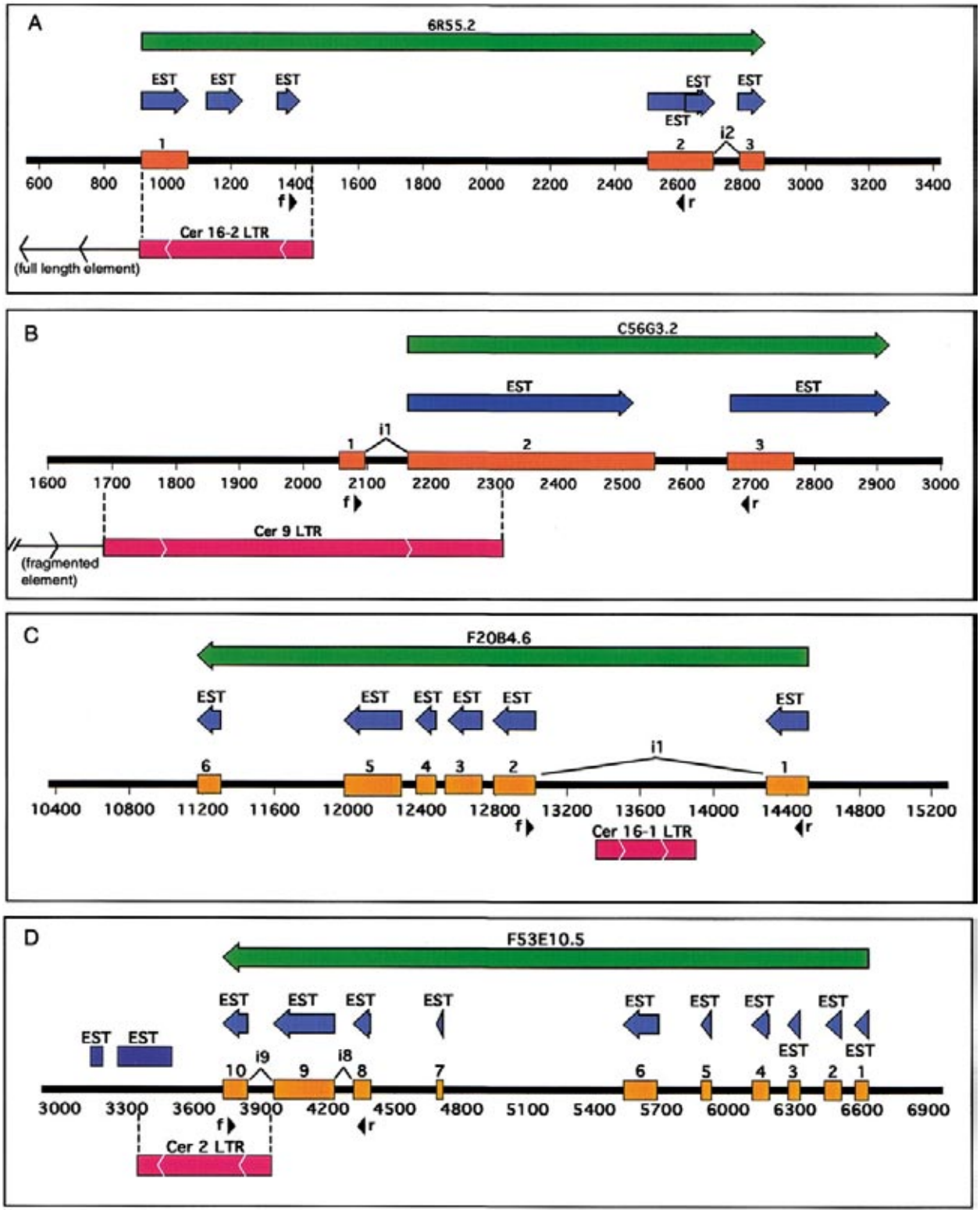
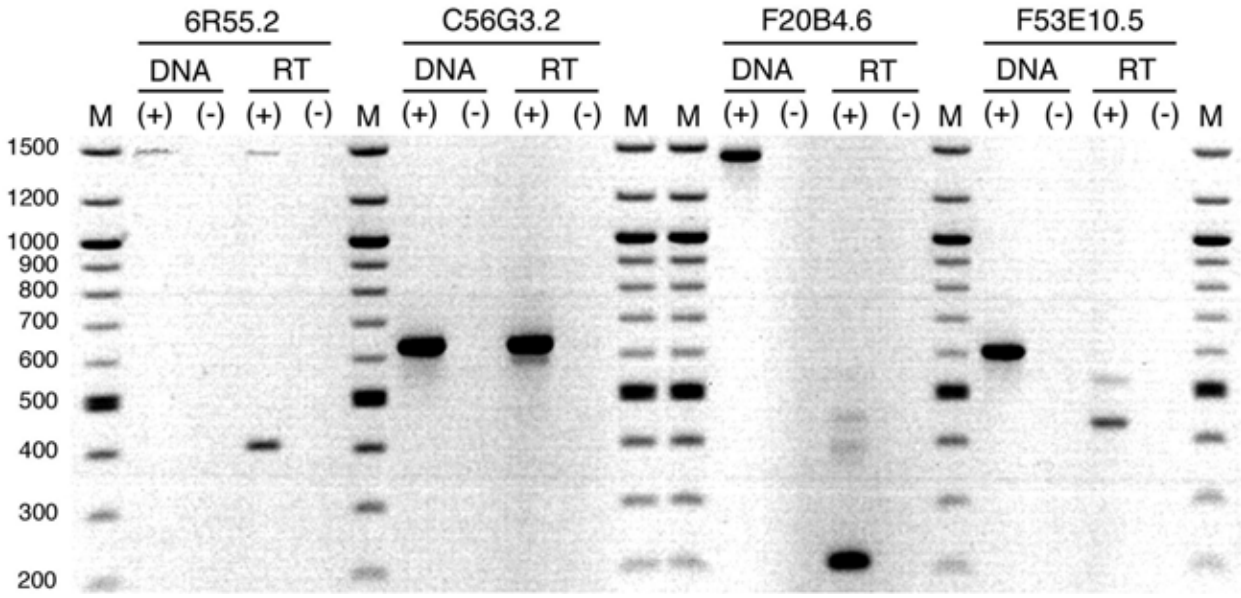

A
6R55.2
EST
EST
EST
EST
EST
EST
1
2
i2
3
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
f
r
Cer 16-2 LTR
(full length element)
B
C56G3.2
EST
EST
1
i1
2
3
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
2600
2700
2800
2900
3000
f
r
Cer 9 LTR
(fragmented element)
C
F20B4.6
EST
EST
EST
EST
EST
EST
6
5
4
3
2
i1
1
10400
10800
11200
11600
12000
12400
12800
13200
13600
14000
14400
14800
15200
f
r
Cer 16-1 LTR
D
F53E10.5
EST
EST
EST
EST
EST
EST
EST
EST
EST
EST
EST
EST
10
i9
9
i8
8
7
6
5
4
3
2
1
3000
3300
3600
3900
4200
4500
4800
5100
5400
5700
6000
6300
6600
6900
f
r
Cer 2 LTR

Figure 2.5. PCR / RT-PCR analysis of *C. elegans* genes containing *Cer* LTR sequence showing the production of spliced, polyadenylated transcripts from these loci.

A negative image is presented for visual clarity. Within a locus, PCR (control) and RT-PCR were performed using the same primer set. DNA (+) and DNA (-) indicate PCR reactions with and without nematode genomic DNA, respectively. RT (+) and RT (-) indicate RT-PCR reactions with and without reverse transcriptase, respectively. M = 100 bp ladder.


6R55.2
C56G3.2
F20B4.6
F53E10.5
DNA
RT
M
(+)
(-)
1500
1200
1000
900
800
700
600
500
400
300
200

CHAPTER 3

EVIDENCE FOR THE CONTRIBUTION OF LTR RETROTRANSPOSONS TO *C. ELEGANS* GENE EVOLUTION[2]

[2]Ganko, E.W., V. Bhattacharjee, P.Schliekelman, and J.F. McDonald. 2003. *Mol. Biol. Evol.*, 20(11):1925-1931. Reprinted here with permission of the publisher.

ABSTRACT

LTR retrotransposons may be important contributors to host gene evolution since they contain regulatory and coding signals. In an effort to assess the possible contribution of LTR retrotransposons to *C. elegans* gene evolution, we searched upstream and downstream of LTR retrotransposon sequences for the presence of predicted genes. Sixty-three percent of LTR retrotransposon sequences (79/124) are located within 1 kb of a gene or within gene boundaries. Most gene-retrotransposon associations were located along the chromosome arms. Our results are consistent with the hypothesis that LTR retrotransposons have contributed to the structural and/or regulatory evolution of genes in *C. elegans*.

INTRODUCTION

The relative abundance of transposable elements (TEs) in eukaryotic genomes varies considerably among species. For example, it is estimated that 3% of the *S. cerevisiae* genome (Kim et al. 1998) and 6% of the *C. elegans* genome (*C. elegans* Sequencing Consortium 1998; Kidwell 2002) are composed of TEs while up to 90% of the genomes of many higher eukaryotes are composed of TEs (e.g., Drosophila,10-20% (Adams et al. 2000; Hoskins et al. 2002; Kaminker et al. 2002); Arabidopsis, 10% (The Arabidopsis Genome Initiative 2000); Mus, 37% (Smit 1999; Waterston et al. 2002); *Homo sapiens,* 43% (Li et al. 2001); Pinus, 90% (Flavell 1986; Pearce et al. 1996). For many years, TEs have been viewed as either neutral or deleterious components of genomes (e.g., Orgel and Crick 1980; Charlesworth, Sniegowski and Stephan 1994). According to this view, TEs located in or near genes ("gene" as used in this paper refers to the transcriptional unit including introns and exons) are likely to be detrimental to gene

function and will be removed by natural selection. Alternatively, TEs can be adaptively beneficial to genes and may contribute to adaptive evolution (e.g., McDonald 1993; McDonald 1995; Brosius 1999; Kidwell and Lisch 2001).

Genomic sequence analysis has proven to be a useful tool in efforts to understand the possible adaptive significance of transposable elements (TEs) in gene and genome evolution. One group of TEs, the retrotransposons, has been studied in this regard. Retrotransposons are the most abundant group of TEs in the human genome and have a lifecycle analogous to that of infectious retroviruses (Boeke et al. 1985). Retrotransposon sequences are transcribed by host transcription complexes, and these transcripts are reverse transcribed by element-encoded reverse transcriptase (RT). As a consequence, retrotransposons contain many *cis*-regulatory components typical of eukaryotic genes, including promoter and enhancer sequences as well as termination and polyadenylation signals (Figure 3.1). The effect of these regulatory sequences are not always limited to the retroelements in which they are contained but may also influence the expression of adjacent genes (e.g., Kapitonov and Jurka 1999; Mager et al. 1999; Baust et al. 2000; Llorens and Marin 2001; Medstrand, Landry and Mager 2001; Stokstad 2001; Jordan et al. 2003). In addition to regulatory effects, retrotransposons may also contribute to the coding regions of genes. For example, in a preliminary study of the human genome, Nekrutenko and Li (2001) discovered that about 4% of human genes have a retrotransposon component in the coding region. Thus, retrotransposons are a significant source of regulatory and coding region variation and a potentially important factor in gene evolution.

To date, whole-genome analyses of the impact of retroelement sequences on gene structure and function has been limited primarily to the human genome (e.g., Nekrutenko and Li 2001; Medstrand, van de Lagemaat and Mager 2002; Jordan et al. 2003). In this paper, we report

the results of a comprehensive genomic study of the contribution of retrotransposon sequences to gene structure and function in the genome of the nematode *C. elegans*. Seventy genes are located within 1 kb of a retrotransposon sequence, i.e., within regions believed to be capable of exerting *cis*-regulatory effects on *C. elegans* gene expression (McGhee and Krause 1997). An additional 40 genes were identified with a retrotransposon within the boundaries of a gene, i.e., in the exons or introns. Further, we show that the observed number of transposons within a 1000 bp-500 bp window of genes is greater than expected by chance. Our results are consistent with the hypothesis that retrotransposons have contributed to the evolution of gene structure and function in *C. elegans*.

RESULTS:

Many retrotransposon sequences are closely associated with genes in *C. elegans*

The search for potential host gene/retrotransposon associations began with a defined dataset of 124 *C. elegans* LTR retrotransposons (Ganko, Fielman and McDonald 2001). One third (1/3) of the retrotransposons in *C. elegans* are *gypsy*-like elements (families *Cer1–Cer6*). The remaining elements are members of the *Bel* clade of retrotransposons (*Cer7–Cer20*). On average, full length *Bel*-like elements are larger than *gypsy*-like elements and their fragments are more numerous in the *C. elegans* genome (Ganko, Fielman and McDonald 2001). Eighty-two *Bel*-like element sequences constitute 356,195 bp (83%) of the retrotransposon component of the *C. elegans* genome while 42 *gypsy*-like elements comprise only 71,728 bp (17%).

The retrotransposon dataset was used to create an annotation file readable by the Wormbase genome browser (Stein et al. 2002). This file was used to visualize the location of

retrotransposons, genes and other genomic features within a given chromosomal region. Analysis of genomic sequence from a 5 kb window on either side of each retrotransposon resulted in the identification of 190 gene/retrotransposon associations (Tables 3.1 & 3.2). Forty (40) retrotransposon sequences were found to be associated with a single gene while 75 were associated with genes both upstream and downstream of the TE. Only 9 retrotransposon sequences were not located within 5 kb of any gene.

Solo LTRs are the most abundant retrotransposon sequence in the *C. elegans* genome (Ganko, Fielman and McDonald 2001), and we found them to be the retrotransposon sequence most frequently associated with genes. However, there was no detectable bias for or against the location of fragments or full-length elements near genes (Table 3.3). The chromosomal distribution of gene-element associations is correlated with the overall distribution of retrotransposon sequences in the genome, i.e., most retrotransposon sequences (Ganko, Fielman and McDonald 2001) and most gene-element associations are located along the chromosome arms (Figure 3.2).

Most *C. elegans cis*-regulatory regions have been shown to extend approximately 1 kb upstream of transcriptional start sites (McGhee and Krause 1997). We find 70 instances where a retrotransposon sequence lies within 1 kb of a gene (Table 3.1, Figure 3.3a & b). The number of retrotransposon sequences located 1 kb upstream of genes is significantly greater than expected ($p<0.025$). Further examination of the 1 kb upstream window revealed that retrotransposon sequences were overrepresented within a 500-1000 bp window ($p<.0029$). This result is significant considering approximately 4% of the *C. elegans* genome is contained within a 500-1000 bp window of intergenic space near genes, while 12% of *Cer* elements are found within the same region.

We also investigated the strand orientation of each retrotransposon in relation to its associated gene. Sense and antisense associations were found to be equally abundant. In 97 associations, the gene and the retrotransposon sequence are both in the sense orientation while in the remaining 93 associations, the gene and retrotransposon are in the antisense configuration. This nearly 50/50 ratio holds for all upstream and downstream associations where the gene and element are located ≤ 4000 bp from one another. For the 9 associations where the gene and element are located > 4000 bp apart, 8 of the 9 associations are in the sense configuration.

Associated gene function and homology

Functional information for each gene associated with a retrotransposon was analyzed in order to confirm the validity of the genes. Several studies have addressed the quality of gene identification and prediction in *C. elegans* (e.g., Harrison, Echols and Gerstein 2001; Reboul et al. 2001; Mounsey, Bauer and Hope 2002). The consensus conclusion of these studies is that 80-90% of *C. elegans*'s predicted genes are "real" or functional, while the remainder are likely pseudogenes or false predictions. We find that 125 of the 190 genes associated with retrotransposon sequences have one or more identifiable functional domains or are members of established homolog families. In addition, about half (49%) of all retrotransposon sequences are associated with genes having medium to high identity with *C. briggsae* homologs as defined by Wormbase (93 *C. briggsae* homologs / 190 total associations). Pooling these findings, we conclude that at least 172 of the 190 genes (90.5%) found to be associated with retrotransposon sequences in our study have functional or phylogenetic support.

Some *Cer* elements are within genes

We discovered 40 genes containing a *Cer* retrotransposon component, meaning a retrotransposon was identified within predicted gene boundaries (hereafter an "internal association"). In some cases a retrotransposon sequence lies within 2 genes, so 35 (of 124) retrotransposons are responsible for the 40 internal associations. Since genic regions represent approximately 52% of the *C. elegans* genome, this result is significantly lower than expected (chi-squared test, expect 64.5 *Cer* TEs, $p<4.28^{-8}$) if insertion sites are assumed to be random.

The frequency of solo LTRs (18), fragments (9) and full-length elements (8) in genes is consistent with the frequency observed for all associations. As with all gene-element associations in *C. elegans*, sense (21) and antisense (19) associations are equally abundant. There are >3X more *Bel*-like (27) than *gypsy*-like (8) element sequences located within the boundaries of genes. This result contrasts with the ≈2X greater number of *Bel*-like element sequences present in the entire *C. elegans* genome. Cer 9 is one *Bel*-like element that accounts for nearly a quarter of all internal associations (Table 3.4).

Thirty-five percent (14/40) of internal associations involve an element exclusively within an intron while 23% (9/40) involve an element exclusively within an exon. Element sequences that extend into both intron and exon regions account for the remaining 43% (17/40) of internal associations. In terms of percent contribution to genes, internally-associated retrotransposons vary from 0.4%–87.3% of the DNA of *C. elegans* genes (Table 3.4). The mean contribution of retrotransposon sequences to internally associated genes (including intron and exon regions) is 23%. Eleven internally associated genes (28%) have EST support, while 18 genes (45%) show homology to *C. briggsae* genomic sequences.

DISCUSSION

C. elegans is an attractive model system for the study of the contribution of TEs to genome evolution. The nematode worm has a tractable, sequenced genome (100 Mb) with an active annotation database (*C. elegans* Sequencing Consortium 1998; Stein et al. 2001). Additionally, the *C. elegans* sister species, *C. briggsae*, is currently being sequenced, and the results will be available for comparative genomics in the near future. Utilizing these resources, along with a data set of *C. elegans* LTR retrotransposons (Ganko, Fielman and McDonald 2001), we have conducted a comprehensive study of the potential contribution of LTR retrotransposons to *C. elegans's* gene evolution.

A total of 124 *Cer* retrotransposon sequences (full-length elements, fragmented elements and solo LTRs) account for 0.4% of the *C. elegans* genome. Searching a 5 kb window both upstream and downstream of each *Cer* element sequence resulted in the identification of 190 gene-retrotransposon associations. Interestingly, 79 (63%) LTR retrotransposons map within 1 kb of a gene. Within this group, we discovered that retrotransposons are overrepresented upstream of genes, specifically in an intergenic region 1000-500 bp from genes. This is significant because most *cis*-regulatory sequences are believed to lie within 1 kb of the transcriptional start site of *C. elegans* genes (McGhee and Krause 1997). An additional 21.1% of all associations involved retrotransposon sequences located within introns, exons, or both.

Reports of TE content in humans indicate that >40% of the genome is composed of retroelement sequences (Li et al. 2001) and an estimated 4% of human protein-coding genes have been found to contain retrotransposon sequences (Nekrutenko and Li 2001). Additional studies suggest that the role of retrotransposon sequences on the regulation of human gene expression may also be significant. For example, it was recently estimated that ~24% of

identified human promoter regions contain retrotransposon sequences (Jordan et al. 2003). Our results indicate that 190 of the 19,000 genes (1.0%) identified in the *C. elegans* genome (*C. elegans* Sequencing Consortium 1998; Reboul et al. 2001) are associated with retrotransposon sequences and that 28% (35/124) of all *Cer* element sequences are located within genes.

In a recent study of the distribution of retrotransposon sequences within the human genome, Medstrand et al. (2002) noted a significant decrease in the density of LTR retrotransposon sequences within 5 kb of genes. Moreover, those retrotransposon sequences located near human genes are relatively recent insertions and most often in an anti-sense configuration with respect to the adjacent gene. The authors interpret these results to suggest that most retrotransposon insertions proximal to human genes, and especially those in a sense configuration, are non-adaptive and selected against. In contrast to the pattern observed in humans, our results demonstrate that well over half of all retrotransposon sequences in the *C. elegans* genome (57.9%) are located in or within 1 kb of genes, with no bias against sense associations observed. At least two hypotheses may help account for these differences.

Protection from deletion or recombination may explain why TEs are close to genes in *C. elegans*. The relatively small size of the *C. elegans* genome has been attributable, in part, to a significantly higher rate of deletion than humans and other animals (Kent and Zahler 2000; Robertson 2000). In addition, *C. elegans* is estimated to have up to a 1440-fold higher rate of genome rearrangement than humans and other mammals (Coghlan and Wolfe 2002). Recombination breakpoints in *C. elegans* are typically associated with repetitive sequences, including retrotransposon sequences (Coghlan and Wolfe 2002). Deletion or recombination events involving retrotransposon sequences in or near genes may have an adverse effect and thus

be selected against. Such a scenario might help explain the clustering of retrotransposon sequences that are not otherwise deleterious in or around genes.

Another possible explanation of the abundance of retrotransposon sequences in or near *C. elegans* genes is that they are of adaptive benefit. Indeed, there is a growing body of evidence from a number of systems (Makalowski 2000; Medstrand, Landry and Mager 2001; Nigumann et al. 2002) that retrotransposon sequences have contributed to adaptive changes in gene structure and regulation.

The central regions of *C. elegans* chromosomes are the general location of "house keeping" genes and other essential genes displaying homology to genes even in distantly related species (*C. elegans* Sequencing Consortium 1998). In contrast, many nematode-specific genes are located along the chromosomal arms. Interestingly, *C. elegans* transposons and other repeats also tend to cluster on the chromosomal arms (Surzycki and Belknap 2000; Ganko, Fielman and McDonald 2001). The chromosomal arms of *C. elegans* are regions of high insertional polymorphism, duplications and intra-chromosomal rearrangements (*C. elegans* Sequencing Consortium 1998). Insertions, duplications, chromosome rearrangements and TEs may all have a role in the evolution of novel genes (Long 2001; Betrán and Long 2002). For these reasons, regions of the chromosomal arms of C. elegans might be viewed as an "evolutionary laboratory" where new genes are created and tested by natural selection. Low mobility species such as *C. elegans* may require a diverse group of specialized genes in order to successfully exploit their environment (Hodgkin 2001), and an ability to rapidly evolve new genes or new regulatory structures may be particularly important to these organisms. The fact that nearly all of the *C. elegans* genes that we have found to be in close association with retrotransposon sequences are located in the chromosome arms suggests that retrotransposon sequences may play a role in the

evolution of new nematode genes. It will be interesting to determine if newly evolved genes in other species, including humans, show a preference for close association with retrotransposon sequences.

METHODS:

Data collection

A flat file annotating the chromosomal position of each retrotransposon was created for use with the Wormbase Genome Browser (www.wormbase.org/db/seq/gbrowse, Stein et al. 2002) using a previously defined dataset of 124 LTR retrotransposons in *C. elegans* (Ganko, Fielman and McDonald 2001). Next, a 5,000 bp-sequence window upstream and downstream of each retrotransposon was visually searched via the Genome Browser for the presence of the nearest predicted gene region. The distance for each association was recorded from the closest retrotransposon coordinate to the nearest open reading frame (ORF) coordinate of the associated gene. In cases where more than one gene was located within 5 kb upstream or downstream of a given retrotransposon, only the most proximal gene on either side was scored as an association. A 5 kb window size was chosen based on estimates of average intergenic distances in the *C. elegans* genome and potential regulatory region size (*C. elegans* Sequencing Consortium 1998).

Information regarding the function, expression and homologs of each gene was collected from various sources. For most genes, information on function and size was available from NCBI and Wormbase gene reports (Spring 2002 data releases). EST data were obtained through BLASTs of the NCBI "est" database (http://www.ncbi.nlm.nih.gov/BLAST/). Exon boundaries were based on reports in Wormbase and NCBI. Conserved domains were predicted with the

NCBI CDD-Conserved Domain Database

(http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). *C. briggsae* homology data was obtained from Wormbase (WABA predictions - Kent and Zahler 2000) and directly from the Washington University *C. briggsae* blast server (http://genome.wustl.edu/projects/cbriggsae, Spring 2002 data). MacVector 7.0 (http://www.accelrys.com/products/macvector/) was used to annotate and collate gene information from all sources as well as provide graphical representations of gene/retrotransposon association regions.

Statistical Analysis

The goal of the statistical analysis was to determine whether the distribution of TEs in the genome deviates from the random expectation, and in particular whether TEs tend to lie near genes. We test two null hypotheses 1) the location of TEs follows a uniform distribution in the non-genic genome (the term "non-genic genome" refers to the non-transcriptional regions upstream and downstream of genes) and 2) the location of TEs follows a uniform distribution throughout the entire genome.

To test the first hypothesis we define windows of length 1000 bp upstream and downstream of each gene. This window is defined to contain only non-genic genome. The window is shortened if the distance to the next gene is less than 1000 bp. A TE is located in the window if its nearest end to the gene is located within the window. The following discussion will be in terms of a window on the 5' end of the gene, but identical arguments apply for the 3' window.

Under the null hypothesis, the probability p that a particular TE is located in an upstream window is simply the length of non-genic genome within 1000 bp of a 5' end divided by the total

length of non-genic genome. Then, the probability that X out of N total TEs is located in upstream windows is given by the binomial distribution.

$$P(X = x) = \binom{N}{x} p^x (1-p)^{N-x}. \quad (1)$$

The use of the binomial distribution assumes that 1) the initial insertion point and subsequent survival of each TE is independent of other TEs and 2) that the probability of insertion in a window and subsequent survival is the same for all TEs. Since the density of TEs relative to the entire genome is low, these assumptions seem reasonable. Given the observed number x of TEs located in the window, the probability under the null hypothesis can be calculated using equation (1).

The quantity p was calculated as

$$p = \frac{\sum_j I_{j,j+1}}{G} \quad (2)$$

Where j is over all genes, G is the total non-genic genome length, and $I_{j,j+1}$ is given by

$$I_{j,j+1} = \begin{cases} \text{distance between genes } j \text{ and } j+1 \text{ if distance } < 1000 \\ 1000 \text{ otherwise} \end{cases} \quad (3)$$

One complication arises when a TE is upstream of two different genes. Equation (2) doesn't consider the orientation of genes and overcounts the amount of genome within the 1000 bp window of the 5' end. In such cases the TE was only counted once, which is conservative.

We also consider a window consisting of non-genic genome that is between 500 and 1000 bp from the nearest 5' end. The new quantity p for this case is found by repeating the calculation in equation (2) with 500 bp instead of 1000 and subtracting from the original p. In order for our calculations to remain conservative, a TE is not counted as being in the window if it

is upstream of two genes. The calculations for the second null hypothesis 2) defined above are carried out in a similar manner, with the obvious modifications.

SUPPLEMENTARY MATERIAL

Please see the lab website (http://www.genetics.uga.edu/retrolab/data.html) for the following:

*Table of associations with retrotransposon name and associated gene ID.

*Representative *Cer* family sequences.

ACKNOWLEDGMENTS

Our laboratory is supported by grants from the National Institutes of Health. We thank Eileen Kraemer (University of Georgia, Computer Science) for statistical advice and helpful discussion.

REFERENCES

Adams, M. D., S. E. Celniker, R. A. Holt et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287**:** 2185-2195.

Baust, C., W. Seifarth, H. Germaier, R. Hehlmann and C. Leib-Mosch. 2000. HERV-K-T47D-Related long terminal repeats mediate polyadenylation of cellular transcripts. Genomics 66**:** 98-103.

Betrán, E., and M. Long. 2002. Expansion of genome coding regions by acquisition of new genes. Genetica 115**:** 65-80.

Boeke, J. D., D. J. Garfinkel, C. A. Styles and G. R. Fink. 1985. Ty elements transpose through an RNA intermediate. Cell 40**:** 491-500.

Brosius, J. 1999. Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica 107**:** 209-238.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282**:** 2012-2018.

Charlesworth, B., P. Sniegowski and W. Stephan. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371**:** 215-220.

Coghlan, A., and K. H. Wolfe. 2002. Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. Genome Res 12**:** 857-867.

Flavell, R. B. 1986. Repetitive DNA and chromosome evolution in plants. Philos Trans R Soc Lond B Biol Sci 312**:** 227-242.

Ganko, E. W., K. T. Fielman and J. F. McDonald. 2001. Evolutionary history of *Cer* elements and their impact on the *C. elegans* genome. Genome Res 11**:** 2066-2074.

Harrison, P. M., N. Echols and M. B. Gerstein. 2001. Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. Nucleic Acids Res 29**:** 818-830.

Hodgkin, J. 2001. What does a worm want with 20,000 genes? Genome Biol 2**:** comment2008.

Hoskins, R. A., C. D. Smith, J. W. Carlson et al. 2002. Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. Genome Biol 3**:** RESEARCH0085-0085.

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

Kaminker, J. S., C. M. Bergman, B. Kronmiller et al. 2002. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol 3**:** RESEARCH0084-0084.

Kapitonov, V. V., and J. Jurka. 1999. The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. J Mol Evol 48**:** 248-251.

Kent, W. J., and A. M. Zahler. 2000. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae-C. elegans* genomic alignment. Genome Res 10**:** 1115-1125.

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

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

Kim, J. M., S. Vanguri, J. D. Boeke, A. Gabriel and D. F. Voytas. 1998. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res 8**:** 464-478.

Li, W. H., Z. Gu, H. Wang and A. Nekrutenko. 2001. Evolutionary analyses of the human genome. Nature 409**:** 847-849.

Llorens, C., and I. Marin. 2001. A mammalian gene evolved from the integrase domain of an LTR retrotransposon. Mol Biol Evol 18**:** 1597-1600.

Long, M. 2001. Evolution of novel genes. Curr Opin Genet Dev 11**:** 673-680.

Mager, D. L., D. G. Hunter, M. Schertzer and J. D. Freeman. 1999. Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (HHLA2 and HHLA3). Genomics 59**:** 255-263.

Makalowski, W. 2000. Genomic scrap yard: how genomes utilize all that junk. Gene 259**:** 61-67.

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

McDonald, J. F. 1995. Transposable elements: possible catalysts of organismic evolution. Trends Ecol. Evol 10**:** 123-126.

McGhee, J. D., and M. W. Krause, 1997 Transcription factors and transcriptional regulation, pp. 147-184 in *C. elegans II*, edited by D. L. Riddle, T. Blumenthal, B. J. Meyer and J. R. Priess. Cold Spring Harbor Laboratory Press, Plainview, N.Y.

Medstrand, P., J. R. Landry and D. L. Mager. 2001. Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J Biol Chem 276**:** 1896-1903.

Medstrand, P., L. N. van de Lagemaat and D. L. Mager. 2002. Retroelement Distributions in the Human Genome: Variations Associated With Age and Proximity to Genes. Genome Res. 12**:** 1483-1495.

Mounsey, A., P. Bauer and I. A. Hope. 2002. Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes. Genome Res 12**:** 770-775.

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

Nigumann, P., K. Redik, K. Matlik and M. Speek. 2002. Many human genes are transcribed from the antisense promoter of l1 retrotransposon. Genomics 79**:** 628-634.

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

Pearce, S. R., G. Harrison, D. Li, J. Heslop-Harrison, A. Kumar and A. J. Flavell. 1996. The Ty1-copia group retrotransposons in Vicia species: copy number, sequence heterogeneity and chromosomal localisation. Mol Gen Genet 250**:** 305-315.

Reboul, J., P. Vaglio, N. Tzellas et al. 2001. Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. Nat Genet 27**:** 332-336.

Robertson, H. M. 2000. The large srh family of chemoreceptor genes in Caenorhabditis nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. Genome Res 10**:** 192-203.

Smit, A. F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9**:** 657-663.

Stein, L., P. Sternberg, R. Durbin, J. Thierry-Mieg and J. Spieth. 2001. WormBase: network access to the genome and biology of *Caenorhabditis elegans*. Nucleic Acids Res 29**:** 82-86.

Stein, L. D., C. Mungall, S. Shu et al. 2002. The Generic Genome Browser: A Building Block for a Model Organism System Database. Genome Res. 12**:** 1599-1610.

Stokstad, E. 2001. Entomology. First light on genetic roots of Bt resistance. Science 293**:** 778.

Surzycki, S. A., and W. R. Belknap. 2000. Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. Proc Natl Acad Sci U S A 97**:** 245-249.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408**:** 796-815.

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

Table 3.1 – Distribution of distances between genes and *Cer* retrotransposons in *C. elegans*

	Element-gene Associations	% total
Internal (within gene)	40	21.1%
1-1000 bp	70	36.8%
1000-2000 bp	33	17.4%
2001-3000 bp	25	13.2%
3001-4000 bp	14	7.4%
4001-5000 bp	8	4.2%
Total	190	

Table 3.2 – Gene / retrotransposon associations per *Cer* family

TE Family	internal (0bp)	external (1-5000bp)	total associations
Cer1	0	7	7
Cer2	2	8	10
Cer3	5	15	20
Cer4	0	0	0
Cer5	0	21	21
Cer6	1	7	8
Cer7	1	2	3
Cer8	2	3	5
Cer9	9	10	19
Cer10	0	6	6
Cer11	0	1	1
Cer12	5	19	24
Cer13	1	2	3
Cer14	0	0	0
Cer15	3	9	12
Cer16	3	21	24
Cer17	1	0	1
Cer19	2	14	16
Cer20	5	5	10
Total	40	150	190

Table 3.3 – Gene / retrotransposon associations for full length, fragmented or solo LTR retrotransposons.

	Full	Frag	LTR	Total
0 associations	1	2	6	9
1 association	7	10	23	40
2 associations	15	16	44	75

Table 3.4 – Genes with a *Cer* retrotransposon component

LTR Family	Gene ID	TE / exon %	TE / intron %	TE / gene %
Cer2	f53e10.5	11.9	5.4	7.8
Cer2-1	k08d10.5	61.7	100.0	62.6
Cer3	f58h7.7	6.0	5.7	5.9
Cer3-1	k09h9.7	0.0	42.9	28.0
Cer3-1	y39b6a.b	0.0	9.6	6.5
Cer3-1	y75b8a.27	0.0	9.0	6.3
Cer3-1	y23h5b.7a	20.8	1.4	3.2
Cer6	y73f8a.11	0.0	4.1	3.2
Cer7	h08m01.2	0.0	2.3	1.8
Cer8	c03a7.12	32.3	76.6	65.4
Cer8	c03a7.13	11.3	0.0	7.7
Cer9	f07b7.14	32.8	39.2	37.0
Cer9	c40a11.1	1.9	0.0	1.5
Cer9	b0047.4	23.4	0.0	21.7
Cer9	f15a2.4	0.0	29.0	19.7
Cer9	f07b7.8	2.3	0.0	1.3
Cer9	k06c4.1	0.8	0.0	0.4
Cer9	c33c12.4	22.8	7.9	12.6
Cer9	c56g3.2	24.8	0.0	19.9
Cer9	y57a10a.30a	15.8	22.5	18.3
Cer12	y60a3a.5	21.4	0.8	3.6
Cer12	w03g1.9	10.4	0.0	4.6
Cer12	c04g6.7	14.6	0.0	6.3
Cer12-1	zc15.5	41.4	46.1	44.6
Cer12-1	zc15.2	4.9	19.7	15.5
Cer13	c09b9.3	0.0	19.4	12.6

Cer15-1	y40h7a.6	36.1	14.1	17.4
Cer15-1	f19b2.1	36.7	82.5	72.9
Cer15-1	f19b2.8	1.3	0.0	0.4
Cer16-1	y71h2am.3	0.0	15.4	11.4
Cer16-1	f20b4.6	0.0	16.0	11.4
Cer16-2	6r55.2	33.6	25.8	27.6
Cer17	r52.10a	0.0	92.3	78.8
Cer19	t06a10.2	86.6	87.5	87.3
Cer19	f35h10.3	11.2	41.3	34.8
Cer20	y87g2a.11	0.0	40.1	35.3
Cer20	f41b5.5	0.0	83.1	69.7
Cer20	r11g10.1b	0.0	14.2	8.5
Cer20	t28d6.2	0.0	12.8	7.7
Cer20	k01d12.3	3.5	54.5	21.8

NOTE: Element / exon % is a function of the number of TE nucleotides within predicted exon boundaries ÷ total number of nucleotides within the exons of a given gene. Element / intron % substitutes the values for intron boundaries. Element / gene % combines the exon and intron calculations.

Figure 3.1: Potential gene/retrotransposon association schemes.

Retrotransposons may provide regulation and/or coding regions to a gene. (**A)** Element acts as enhancer of host gene. (**B**) Element acts as polyadenylation or promoter signal within host gene. (**C**) Element contributes exon material. Schemes are not exclusive. For instance, a TE could provide both promotion and exon material.

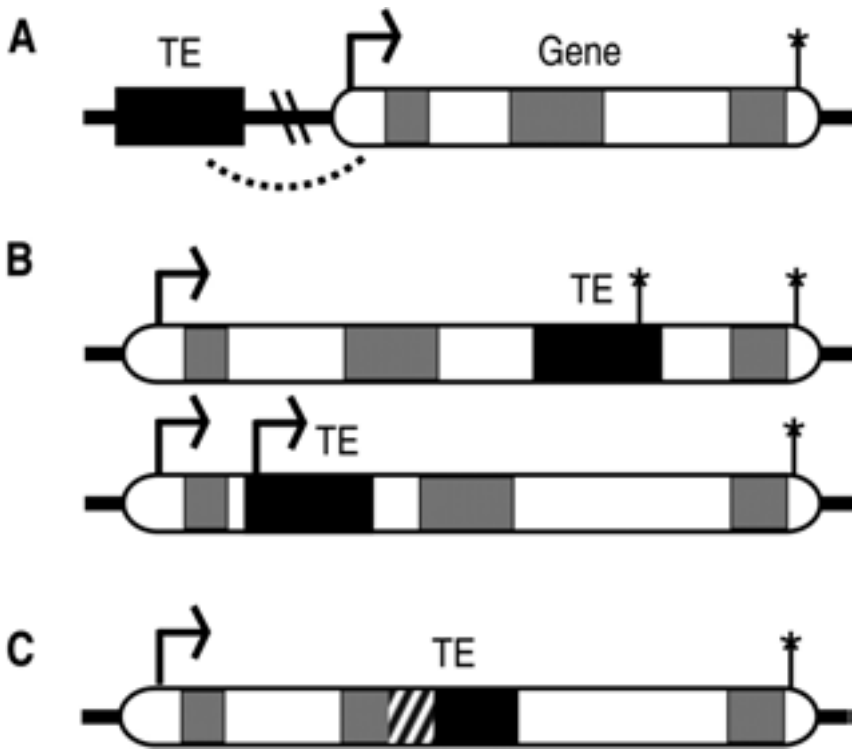

A
TE
Gene
B
TE
TE
C
TE

Figure 3.2: Distribution of gene/retrotransposon associations in the *C. elegans* genome.

A genomic coordinate value for each *Cer* retrotransposon was calculated and plotted to its previous chromosome location (see Ganko, Fielman and McDonald 2001). Chromosomes were divided into three regions (left, centric, right) marked by vertical hash marks. Open circles represent retrotransposons with an associated gene while each retrotransposon located inside a gene is marked by a closed circle. Retrotransposons lacking a gene within 5 kb are marked by an *x*.

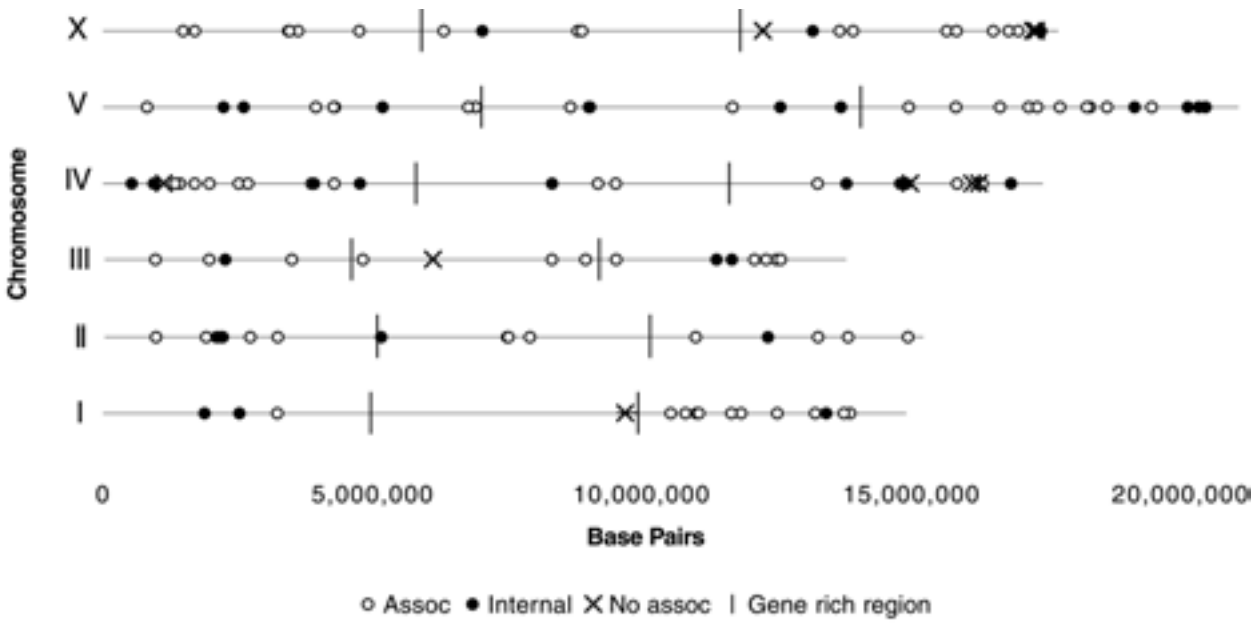
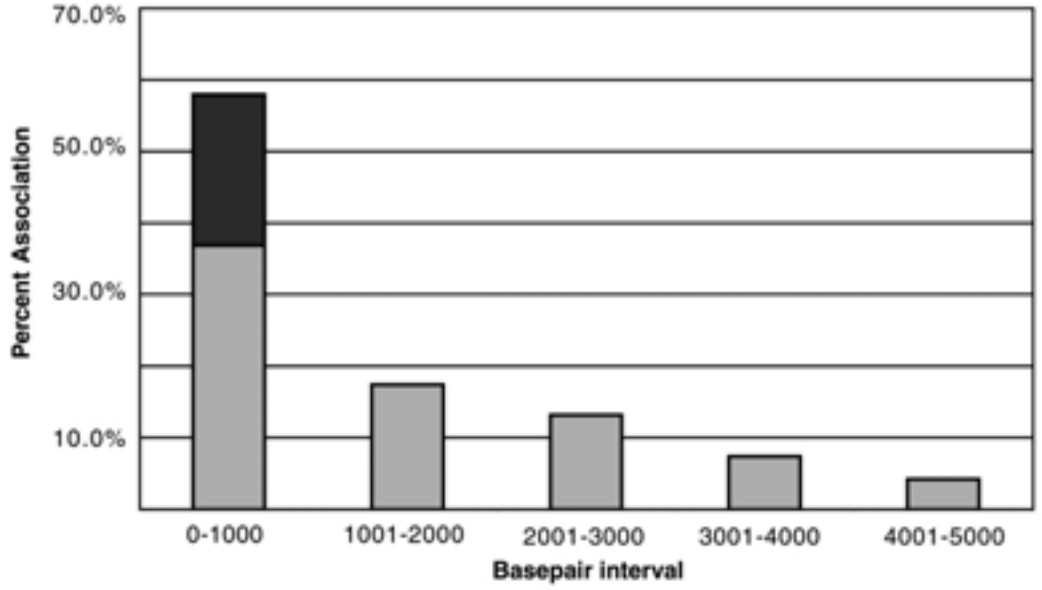

X
V
IV
III
II
I
Chromosome
0
5,000,000
10,000,000
15,000,000
20,000,000
Base Pairs
Assoc
Internal
No assoc
Gene rich region

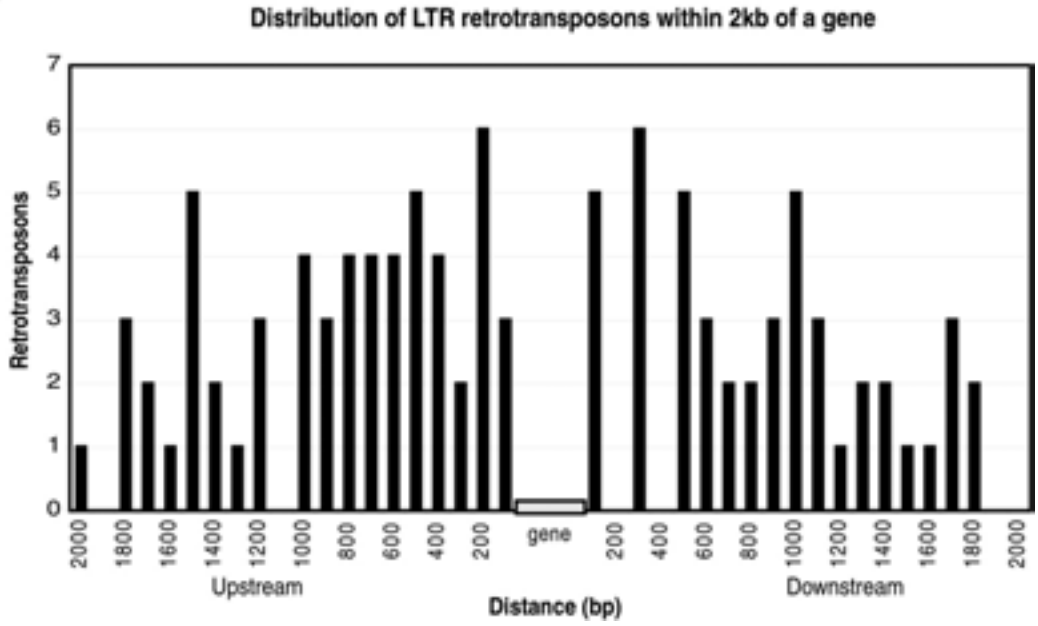
Figure 3.3: Distance distributions between LTR retrotransposon and associated gene.

(**A**) 190 gene/LTR retrotransposon associations within a 5 kb window were sorted into 1 kb bins. Dark gray shading denotes internal element contribution (top rectangle of the 0-1,000 bp column). (**B**) Distribution of retrotransposon associations upstream and downstream of a gene within a window of 1 bp to 2000 bp. A model gene is represented by the small gray bar in the center and is not scaled to size.

A


Percent associations per 1kb distance intervals
70.0%
50.0%
30.0%
10.0%
Percent Association
0-1000
1001-2000
2001-3000
3001-4000
4001-5000
Basepair interval

B


Distribution of LTR retrotransposons within 2kb of a gene
Retrotransposons
0
1
2
3
4
5
6
7
2000
1800
1600
1400
1200
1000
800
600
400
200
gene
200
400
600
800
1000
1200
1400
1600
1800
2000
Upstream
Downstream
Distance (bp)

CHAPTER 4

LTR RETROTRANSPOSON-GENE ASSOCIATIONS IN *DROSOPHILA MELANOGASTER*[3]

[3]Ganko, E.W., C. Greene, J.A. Lewis, V. Bhattacharjee, and J.F. McDonald. To be submitted to Genome Research.

ABSTRACT

LTR retrotransposons are common to many genomes, and due to incorporated regulatory and coding signals, may contribute to gene evolution. To better understand the potential contribution of LTR retrotransposons to *D. melanogaster* gene evolution, we searched upstream and downstream of each identified LTR retrotransposon sequence for the presence of a neighboring gene. A total of 228 (33% of 682) LTR retrotransposon sequences were found to be located in or within 1,000 bp of a gene. Full-length and near full-length LTR retrotransposons are significantly more likely to be located in or within genes than are small, fragmented LTR retrotransposons. Genes containing an LTR retrotransposon sequence present within their boundaries are >5x larger than the size of the average *D. melanogaster* gene. Genes encoding signal transduction, behavioral, and developmental functions are preferentially associated with LTR retrotransposon sequences, while genes encoding physiological processes tend not to be associated with LTR retrotransposon sequences. Taken together, these results support the hypothesis that LTR retrotransposons may contribute to gene evolution in *D. melanogaster*.

INTRODUCTION

Transposable elements (TEs) are mobile sequences abundant within eukaryotic genomes (e.g., *Drosophila melanogaster*, 10-20% (Hoskins et al. 2002; Kaminker et al. 2002)*; Homo sapiens*, >40% (Li et al. 2001); Lillium, >90% (Leeton and Smyth 1993)). Although TEs can be maintained in populations on a day-to-day basis even in the face of slight negative selection (e.g. Doolittle and Sapienza 1980; Orgel and Crick 1980;

Charlesworth, Langley and Sniegowski 1997), this does not preclude the possibility that TE sequences may contribute significantly to gene and genome evolution over evolutionary time (e.g. McDonald 1993; McDonald 1995; Brosius 1999). Indeed, there are now many examples of TE sequences having contributed significantly to gene and genome evolution in a variety of species (e.g. Makalowski 2000; Medstrand, Landry and Mager 2001). With the availability of sequence databases for a number of species, it has become possible to conduct systematic genome searches for TE-gene associations in order to objectively assess the potential contribution of these elements to gene evolution. For example, recent analyses in the human genome have shown that retrotransposon sequences are present in the intron and exon regions of ~4% of genes (Nekrutenko and Li 2001), in the untranslated regions (UTR) of ~27% of genes (van de Lagemaat et al. 2003), and in ~25% of promoter regions (Jordan et al. 2003). We recently reported that LTR retrotransposon sequences are present within the regulatory region and/or the transcription boundaries in 0.6% of *C. elegans*' genes (Ganko et al. 2003). In this paper, we report the results of a detailed analysis of the association of long terminal repeat retrotransposon (hereafter, LTEs) sequences with genes in the *Drosophila melanogaster* genome.

LTEs are a class of transposable elements that have a lifecycle analogous to that of infectious retroviruses (Boeke et al. 1985). LTEs are initially transcribed into RNA by the host organism's transcriptional machinery, and subsequently reverse transcribed by element-encoded reverse transcriptase (RT) to create a DNA copy. In order to initiate RNA transcription, retrotransposons contain *cis*-regulatory sequences typical of eukaryotic genes, including promoter, enhancer and termination signals. The regulatory

effects of LTE signals are not limited to the retroelements in which they are contained, and may influence the expression of adjacent genes. In addition, LTE sequences may also be incorporated into the coding regions of genes. Thus, LTE insertions that do not destroy genetic functionality may be a potential source of adaptive genetic variation (e.g. McDonald 1993; McDonald 1995; Brosius 1999; Makalowski 2000).

Drosophila melanogaster is good model for evolutionary genomic studies because of the availability of a high-quality genome sequence (Adams et al. 2000; Celniker et al. 2002) and annotation (Misra et al. 2002), especially with regard to transposon sequences (Kaminker et al. 2002). We report here the identification and preliminary characterization of 82 LTE sequences located within 1 kb of a gene and an additional 146 LTEs located inside gene boundaries. Genes with LTE sequences located within their boundaries are significantly larger (~5x) than the average *D. melanogaster* gene. LTE sequences are preferentially associated with recently evolved genes encoding signal transduction, behavioral, and developmental functions. Our results are consistent with the hypothesis that LTEs are a significant contributing factor to *D. melanogaster* gene evolution.

RESULTS

One-third of all identified LTR retrotransposon sequences are located in or within 1,000 bp of a gene

The *Drosophila melanogaster* genome is estimated to contain 14,000 genes (Adams et al. 2000; Misra et al. 2002). Recently, 682 full-length and/or partial LTE

sequences (partial sequences are defined as solo LTRs, truncated LTRs and/or truncated full-length elements) have been identified in the euchromatic portion of the *Drosophila* genome (Kaminker et al. 2002). Perl scripts were developed to determine the distance from each of these 682 LTE sequences to the nearest flanking genes. Since most sequences known to exert *cis*-regulatory effects on *Drosophila* gene expression are located within 1,000 bp of the transcriptional start site (Papatsenko et al. 2002), we limited our dataset to genes with LTE sequences located within 1,000 bp upstream or downstream of established genes or within gene boundaries (introns or exons). This dataset contains LTE-gene associations of potential adaptive significance.

Our results (Table 4.1) indicate that 228 or 33.4% of LTR retrotransposon sequences located in the euchromatic region of the *Drosophila* genome are associated with genes. There are 82 LTE sequences located 1 kb upstream or downstream of 102 genes (proximal associations) and 146 LTE sequences located within the introns or exons of genes (internal associations). Proximal associations are comprised of element sequences distributed equally upstream and downstream of genes (53/102 upstream; 49/102 downstream). Likewise, there is no significant bias in the sense orientation of element sequences located proximal to genes (upstream: 30/53 elements in sense orientation with respect to the associated gene; downstream: 23/49 in sense orientation; $p>0.10$). This result contrasts with the results of a recent study of the human genome where it was found that retrotransposon sequences located in human genes are most often in an anti-sense configuration with respect to the adjacent gene (Medstrand, van de Lagemaat and Mager 2002), while retrotransposons in the 5' and 3' untranslated regions are significantly more likely to be in a sense configuration (van de Lagemaat et al. 2003).

However, our results in *Drosophila* are similar to the relatively equal sense/antisense distributions of LTEs near genes observed in *C. elegans* (Ganko et al. 2003).

In order to determine whether the observed number of LTE sequences associated with genes is greater or less than what is expected by chance, we computed an expected number of associations based on the probability of an insertion event occurring randomly within a 1-1000 bp proximal window or within the transcriptional boundaries of any annotated gene in the genome. The observed number of proximal associations (obs: 82) is not significantly different from what is expected by chance (exp: 85; $p>0.10$). In contrast, the observed number of internal associations (obs: 146) is significantly less than what is expected by chance (exp: 382; $p<0.001$), presumably due to negative selection.

Consistent with a random distribution model, we found that, as a general rule, those LTE sequences that are most abundant in the genome are also the sequences most frequently associated with genes. There were, however, some notable exceptions. For example, families *DM88, GATE, invader 1* and *invader 3* have significantly fewer LTE-gene associations than expected (χ^2 =72.3, d.f.=43, p=0.003, Table 4.1) based on the number of family members in the genome. Intra-family transposon clustering is the likely cause of the low percentage of associations in all four of these families. For example, 30 of the 32 *DM88* elements are located within a 32 kb stretch of chromosome 3R, and 18 of 26 *invader1* elements are located along a separate 28 kb stretch of chromosome 3R. Only LTE sequences on the edge of a cluster can be near genes, so transposon clustering generally prevents LTEs from associating with host genes and may explain the limited associations of *DM88, GATE, invader 1* and *invader 3* elements.

The distribution of LTE-gene associations is not correlated with gene density

While the accumulation of most LTEs does not appear to be tightly correlated with regional gene density (Bartolome, Maside and Charlesworth 2002; Rizzon et al. 2002), it remains a possibility that those LTEs associated with genes may lie within chromosomal regions of high gene density. To test this possibility we determined gene densities across consecutive 200 kb regions of each chromosome. The mean number of genes in each bin was calculated for all regions of the chromosome, then for regions that contained at least one LTE, and finally for regions that contained an LTE-gene association. Neither LTEs nor LTE-gene associations accumulate in regions significantly more dense than the mean gene density of the individual chromosome (Table 4.2).

To test if LTE associations were more likely to occur between genes with small intergenic distances, we measured the distance from each gene to its neighbor. The overall mean distance from gene to gene (4,483 ± 638 bp) is essentially the same as the distance between genes with a proximal LTE (4,324 ± 1,366 bp, disregarding the LTE sequence). Thus, neither the regional or local density of genes is a good predictor of LTE-gene associations.

Most LTEs located in or proximal to genes are full-length or near full-length in size

Most *D. melanogaster* full-length retrotransposons are relatively recent insertions (Bowen and McDonald 2001; Kaminker et al. 2002; Lerat, Rizzon and Biemont 2003). Our results indicate that LTEs associated with genes are significantly larger (5765 ± 178 bp) than the size of the average *Drosophila* LTE sequence (4531 ± 242 bp) in the *D. melanogaster* genome. This suggests that most LTE-gene associations are of recent evolutionary origin.

To further investigate whether recent insertions were disproportionately more likely to be associated with a gene rather than older insertions, we looked at the size distribution of all LTEs. Using a representative consensus sequence from each LTE family as the expected reference size of a full-length element, we found that 348 LTE sequences are ≥90% of the consensus size (near full-length). Another 123 LTE fragments range from 21% - 89% of consensus size (medium), and the remaining 211 LTEs are ≤20% of consensus size (small), consisting of 153 fragments and 58 solo LTRs. When the size distribution of all LTEs is compared to the size distribution of LTEs associated with genes (Figure 4.1), small LTEs were found to be consistently underrepresented (obs 30, exp 67) while large LTEs are were found to be associated with genes more frequently (obs 153, exp 112) than expected based on a random model of association (χ^2=47.09, d.f.=20, p=0.0006).

The LTE size data was further analyzed to determine if the LTE size groups (full-length/near full-length, medium sized, small sized) were equally distributed both within genes and proximal to genes. While the ratio of medium sized LTEs varies little within the proximal or internal association groups, the ratio of small LTEs proximal to and within genes decreases, and the ratio of full-length/near full-length fragments increases (Figure 4.2). Small LTEs comprise 31.0% of all LTEs in the genome but only account for 17.1% of LTE-gene proximal associations In contrast, full-length/near full-length LTEs comprise 49.5% of all LTEs in the genome but account for 64.6% of all LTE sequences associated with genes. For LTEs inside genes, small LTEs are even less frequent, accounting for only 11.0% (16 / 146) of LTE sequences located within gene boundaries compared to full-length/near full-length sequences which account for 69.2%

of LTEs internally associated with genes (101 / 146 total). Thus, recently inserted (full-length/near full-length) elements account for the majority of LTEs associated with genes.

As a general rule, genes involved in basic cellular functions are relatively conserved across taxa while more recently evolved, more specialized genes are taxa specific (e.g. van de Lagemaat et al. 2003; Castillo-Davis et al. 2004). In order to determine if LTE sequences are differentially associated with these different classes of genes, we analyzed the pattern of LTE associations with *Drosophila* genes that have homologues across a broad spectrum of species. Utilizing the 2,503 *Drosophila* genes represented in the NCBI-curated homologene dataset of putative orthologous genes (http://www.ncbi.nlm.nih.gov/HomoloGene/), we identified 51 genes that were either internally or proximally associated with an LTE sequence. We found that 64.7% (33 / 51) of homologenes were associated with full-length/near full-length LTE sequences which is not significantly different from the overall frequency of LTE-gene associations involving this size class. In contrast, only 5.9% (3 / 51) of homologenes were associated with a small, presumably older, LTE sequences. This value is significantly less than the frequency of small LTE-gene associations overall. Thus, while newly inserted LTEs (i.e. full-length/near full-length LTEs) appear to insert in or near all classes of genes with equal frequency, over time, LTE-homologene associations are apparently being selected against.

Genes associated with LTEs are larger than average

Variation in the size of *Drosophila* genes is due in part to variation in the size of introns. LTE insertions into large genes might be less likely to disrupt coding regions and thus less likely to be eliminated by natural selection. To test if LTEs are

preferentially associated with large genes, we compared the mean size of all genes to the mean size of associated genes (excluding the size of the inserted LTE sequence). The results indicate that genes associated with an LTE are nearly 5X larger than the average size of *Drosophila* genes (Figure 4.3). Homologenes with an LTE association follow a similar trend. When grouped by associating LTE size, genes associated with full-length/near-full length LTEs are again nearly 5X larger than the average gene. Genes associated with small LTEs display tremendous size variation, but on average are again larger than the typical *Drosophila* gene (Figure 4.3). Interestingly, while the size of genes with internal LTE associations are also larger than average, those genes with proximal LTE associations are of average size.

We found that introns in genes with an internal LTE are more numerous and significantly larger than the average sized intron (Table 4.3). While exons are more numerous in genes with an internal LTE, they are not significantly larger than average sized exons (Table 4.3). These results are consistent with the hypothesis that larger genes (with larger/more numerous introns) are more tolerant of LTEs. In contrast, the effect of LTE insertions proximal to genes does not seem to be affected by the size of the associated gene.

Large LTEs are preferentially associated with several functional categories of genes

Several authors have noted that transposons are preferentially associated with certain functional classes of genes (e.g. Grover et al. 2003; van de Lagemaat et al. 2003). To investigate this question in *Drosophila*, we grouped our LTE-gene associations using gene ontology (GO) terms. GO terms are descriptors of gene product characteristics hierarchically categorized under three root terms ('cellular component', 'biological

process', and 'molecular function'). Using a custom set of Perl scripts, we traced each *Drosophila* gene descriptor to its respective root term. The cumulative results for all *Drosophila* genes were used to calculate expectation values for the descriptors of our subset of LTE associated genes. For full-length/near full-length LTEs, no significant differences were observed in the cellular component or molecular function (Table 4.4) however, the biological process group displayed significant deviation from the random expectation (χ^2 p=8.1E-25, Table 4.5).

Individual analysis of biological process terms (Table 4.5) demonstrated that the subordinate descriptors 'development' (obs 225, exp 166, p=1.4E-07) and 'behavior' (obs 32, exp 9, p=1.4E-09) were overrepresented, while the 'physiological processes' descriptor was underrepresented (obs 255, exp 329, p= 2.5E-09). The subset of homologenes that are associated with LTEs display a similar pattern to that of associated genes (development obs 112/ exp 70, behavior 12/4; physiological processes 105/149).

We further analyzed the subordinate descriptor terms of the three significant biological processes (Table 4.6). Significant deviation from expectation was not observed among individual descriptors of the behavior group, though 'learning and/or memory' (obs 10, exp 5) was twice the expected value. Two development descriptor terms were significantly different. 'Pattern specification', defined as patterns of cell differentiation, was underrepresented (obs 8, exp 32, p=7.0E-08), while 'morphogenesis' was overrepresented (obs 115, exp 92, p=1.2E-03). The subordinate descriptor term 'morphogenesis of an epithelium' (obs 13, exp 4, p=0.002) was the lone significantly overrepresented morphogenesis term. Two physiological process descriptor terms were also significantly different than expectation. 'Metabolism' was underrepresented (obs 94

, exp 131, p=1.2E-06) while 'response to external stimulus' was overrepresented (obs 69, exp 37, p=1.1E-07). Taken together, large LTEs in *Drosophila* appear to preferentially associate with genes in select functional groups, including morphogenesis of an epithelium, response to external stimulus, and behavioral functions, while associations with genes involved in metabolism and patterns of cell differentiation are significantly fewer than expected.

Small LTEs are preferentially associated with signal transduction genes

Only the molecular function group displayed significant differences within the small LTE association dataset ($\chi 2$ p=9.5E-19, Table 4.4) and a binomial analysis confirmed that 'signal transduction' was overrepresented among small LTE associations (obs 19, exp 3, p=1.6E-11). A greater than expected number of signal transduction terms within LTE associated homologenes was also observed (obs 14, exp 4, p=.018). This is especially remarkable since signal transduction is underrepresented in the whole homologene set (obs 194, exp 307). 'Receptor activity', 'receptor binding', and 'receptor signal protein activity' are the descriptor terms subordinate to signal transduction in the GO hierarchy, though only receptor activity is overrepresented (obs 18, exp 14, p=.046) among small LTE associations. No subordinate terms of receptor activity were significantly different than expectation.

DISCUSSION

Drosophila melanogaster has a relatively small genome with approximately 120 Mbp of euchromatin (Adams et al. 2000) containing 682 LTEs (Kaminker et al. 2002).

We have investigated the potential contribution of LTEs to *D. melanogaster* gene evolution by analyzing all 682 euchromatic LTEs for the presence of nearby genes. We found that 248 of the 13,300 genes (1.9%) identified in the *D. melanogaster* genome have LTE sequences proximal to or within transcription boundaries, and that 21.4% (146 / 682) of all LTE element sequences are located within genes.

Compared to the 100Mb genome of the nematode *C. elegans*, *Drosophila* has fewer genes (19,000 vs. 13,600 respectively), though a roughly equal number of transcripts (18,000 - Reboul et al. 2001; Misra et al. 2002). *C. elegans* has 124 LTEs in its genome, most fragments (Ganko, Fielman and McDonald 2001), and features a higher percentage (64%) of LTEs within or proximal to genes. Of the 124 LTEs, 79 have inserted within or proximal to 110 genes (0.6% of genes), and it is believed that many of these associated genes have species-specific functions (Ganko et al. 2003).

Though our results in *Drosophila* are comparable to those from the small *C. elegans* genome, they contrast with findings from studies of the much larger human genome (~3.2 Bbp). In the human genome, >40% of the genome is composed of retroelement sequences (Li et al. 2001) and approximately 4% of protein-coding genes are believed to contain retrotransposon sequences (Nekrutenko and Li 2001). Indeed, a growing body of evidence in humans indicates that retrotransposon sequences have contributed to adaptive changes in gene structure (e.g. Makalowski 2000; Medstrand, van de Lagemaat and Mager 2002; Nigumann et al. 2002). Additional studies suggest that retrotransposon sequences may significantly influence the regulation of human gene expression. For example, it was recently estimated that ~24% of identified human promoter regions (Jordan et al. 2003) and 27% of genes with an identified untranslated

region (van de Lagemaat et al. 2003) contain retrotransposon sequences. These prior results in humans display a remarkable potential for transposon contribution to gene structure and regulation, and contrast with the smaller potential observed in *Drosophila* and *C. elegans*. The number of transposon sequences is a key difference between small genome organisms such as *D. melanogaster* and *C. elegans* and larger genome organisms such as mouse and human. Organisms with small genomes may utilize genome defense strategies that prevent the accumulation of transposons, thereby limiting the number of transposon–gene associations.

The adaptive potential of LTEs in Drosophila

The high rate of DNA deletion in the *Drosophila* genome has been implicated as a transposon removal process (Moriyama, Petrov and Hartl 1998, Petrov, 2002 #53; Petrov and Hartl 1998). Over evolutionary time it is thought that full-length elements are generally disrupted and removed due to deletion, recombination, and possibly other genomic mechanisms. Compared to a random insertion model we discovered fewer internal LTE-gene associations than expected, likely indicating the loss of older LTE sequences from genes via these removal processes. The skewed ratio of many large, recently inserted LTEs and fewer small, old LTEs inside conserved homologenes indicates that most of the insertions are relatively recent. Furthermore, old elements that have managed to remain are inside comparatively less conserved genes. This result is comparable to previous work on transposon-gene associations in humans and mouse where young, species-specific type genes (the opposite of homologenes) are more likely to have a TE component in humans and mouse (van de Lagemaat et al. 2003). In general,

transposons are more likely to contribute an adaptive function to genes that are rapidly evolving.

Though *Drosophila* may have a lower number of transposons with adaptive potential compared to organisms with large genomes, recent work has provided evidence for the adaptive significance of some transposon fragments in the *Drosophila* genome (Maside, Bartolome and Charlesworth 2002; Franchini, Ganko and McDonald 2004; Schlenke and Begun 2004). For example, Franchini et al (2004) demonstrated that while near full-length elements tested among populations of *D. melanogaster* from around the world are found only in the sequenced strain or in limited populations, two fragmented LTEs were found in all populations tested. One element had putative functionality evidenced by a reduced mutation rate compared to the nearby coding region (Franchini, Ganko and McDonald 2004). Transposon fragments of functional significance have also been reported in *D. melanogaster* for a Hsp70 gene (Maside, Bartolome and Charlesworth 2002) and the insecticide resistant gene Cyp6g1 (Daborn et al. 2002). Additionally, many fragmented LTE-gene associations have been identified in *D. melanogaster* heterochromatic regions and form a substantial component of these genes (Dimitri, Junakovic and Arca 2003). This growing body of evidence points to the possibility that older, fragmented transposons may be important adaptive targets. Over time, most transposons will be removed and/or mutated beyond recognition. Even transposons that provide functionality are not likely to remain whole forever; rather, mutation and deletion mechanisms may act on the nonfunctional portion of the sequence and reduce the LTE to a functional core. This process does not exclude recent insertions

from providing adaptive benefit, but predicts that these insertions are likely to shrink over time.

Our results lead us to believe there are two distinct classes of LTE-gene associations in *Drosophila*. Most large LTEs are likely recent insertions that have inserted in a nearly random fashion around the genome, upon which genome deletion processes have not yet acted. This near-random insertion pattern helps explain their prevalence in regions in and around conserved homologenes. A similar phenomenon among young HERV elements has been described in humans (Medstrand, van de Lagemaat and Mager 2002). We would expect the sequence of a recent insertion with adaptive significance to be reduced to a functional core over time. The second class of LTE-gene associations are the small, and likely more ancient, LTE insertions. Due to age, most small LTEs that are deleterious have been affected by selection mechanisms including deletion and recombination, leading us to find small LTEs in and around genes at a rate lower than random insertion models predict. The low percentage of small LTEs associated to conserved homologenes is further evidence of this trend, as homologenes are more likely to have critical biological importance. Small elements that remain are less likely to have a negative effect, and may even have adaptive significance. Though already reduced, small LTEs with adaptive significance may be further reduced to a functional core over time.

Genes associated with LTEs have functions consistent with response-type genes

Patterns of gene function have been discovered in human genes with associated transposons. For instance, human Alu elements are likely to be associated with genes having metabolism, transport and signaling functions (Grover et al. 2003). In contrast,

human and mouse LTEs have been shown to avoid highly conserved genes (e.g. metabolism, enzymatic domains, replication) and accumulate in the untranslated regions of fast-evolving response genes (e.g. defense, stress, external stimuli) (van de Lagemaat et al. 2003). Even though full-size and near full-size LTEs in *Drosophila* likely have a different insertion history than small LTEs, and their associated genes have distinct functional profiles, both groups associate with external response-type gene functions.

Full-size and near full-size LTEs are significantly underrepresented near genes with functions including pattern specification (of cell differentiation) and metabolism. Genes with these functions are more likely to have a critical biological role and corresponding lower tolerance of disruption. Disruptive insertions into or near genes with critical functions are likely to significantly impair the host and stimulate strong negative selection. Near full-size LTEs are significantly abundant near response-type genes with functions including morphogenesis of an epithelium, response to external stimulus, and behavior. These three functional categories are not critical biological functions and insertions near them may be less disruptive to the survival of the host. Besides the disruption hypothesis, it is possible that recent LTE insertions are not targeting gene types, but are simply inserting into these regions due to easier integration. For example, genes in these functional groups may have relaxed chromatin or transcribe at a time point that is conducive to new LTE integrations. Still, we do not see any clear trends between tissue expression and genes with a large LTE association.

Small LTEs are significantly abundant near genes with receptor activity function. Associations of small LTEs to receptor genes may be abundant for non-adaptive reasons, i.e. these genes may have been readily available integration sites in the past. Another

possibility is that further deletions around the insertion are disruptive to gene function making these genes a "safe harbor" against further deletion. Finally, it is possible that small LTEs provide an adaptive function to genes with receptor activity function. At a basic level, receptors allow a cell to interact with and respond to other cells and the environment. An associated LTE could provide *cis*-regulatory mechanisms (e.g. enhancer, promoter) that change the temporal or physical transcription of receptor genes, thus providing an adaptive change. Molecular tests and sequencing of these regions in combination with population work will be necessary to validate these hypotheses.

METHODS

LTE-gene association data

Annotated chromosome files (Release 3.1) were downloaded from the Berkeley Drosophila Genome Project website (ftp://ftp.fruitfly.org/pub/download/dmel_RELEASE3-1/FASTA/) in Spring 2003. The distance from each annotated LTE (Kaminker et al. 2002) to the closest flanking gene on each side of the LTE was determined, with the exception of the centromere and telomere termini where a transposon may have only one flanking gene. We filtered these results by defining an LTE-gene association as an LTE ≤1000 bp of a gene, based on earlier findings that *D. melanogaster cis*-regulatory regions may extend approximately 1 kb past the transcriptional boundaries (Papatsenko et al. 2002). Thus, all LTEs included in our analyses were either in gene boundaries or within 1000 bp of a gene (Table 4.1). We define "internal association" as an LTE inside the defined transcription borders of a gene,

and define "proximal association" as an LTE within 1-1000 bp of gene boundaries. Expectation values for associations in Table 4.1 were determined using the distribution ratio of each LTE family, that is, the number of individuals in a given LTE family divided by the sum of all identified LTEs. This family distribution value was multiplied by the sum of all associations, the sum of internal LTE-gene associations, or the sum of proximal associations to provide an expectation for a given LTE family in the respective category.

Information regarding the function, size, chromosomal position, gene ontology, and expression of each gene was collected primarily from Flybase (http://flybase.bio.indiana.edu/) gene reports (Spring 2003 data releases). Gene size was determined using the most distant start and stop nucleotides in the case of multiple transcripts. Homologous gene data was obtained from the Homologene database (http://www.ncbi.nlm.nih.gov/HomoloGene/), an NCBI-curated dataset of putative orthologous genes between important model organisms (Zhang et al. 2000). Tests of a distribution model for LTEs internal and proximal to genes were carried out by binomial tests as reported previously (Ganko et al. 2003).

Size and density analyses

To measure gene density, each chromosome was divided into successive 200 kb regions, and the number of genes in each region summed. The gene density of each bin was calculated for the entire chromosome, then for all regions that contained at least one LTE, and finally for regions that contained an LTE-gene association. As a second measure of gene density, we compared the mean intergenic distance between all genes to the intergenic distance of genes with an associated LTE.

Consensus sizes for individual families were determined from Flybase (Kaminker et al. 2002) or RepBase (Jurka 2000), and the size of each individual LTE element was compared to the size of the consensus sequence for the appropriate LTE family to calculate the "percent consensus size." The results were separated into three categories: near full-length (LTEs ≥90% of the consensus size), medium sized (21% - 89% of consensus size), and small sized (LTEs ≤20% of consensus size). Expectation values were calculated based on the ratio of LTEs in a given size bin to all LTEs in the genome.

Functional analysis of genes

Genes were classified into functional categories based on Gene Ontology (GO) terms. The Gene Ontology project has created a controlled vocabulary describing the functional products of genes (Ashburner et al. 2000; Harris et al. 2004). To investigate this defined hierarchical classification we created a set of Perl scripts to trace genes from a specific GO ID to the general descriptors. For example, the ID GO0004871 has a specific description of "signal transducer activity" as a general "molecular function". Performing a trace on a set of genes results in a functional profile that can then be compared to the functional profile of other gene sets. Chi-square tests were used for initial profile comparisons, followed by binomial tests on individual descriptor terms. We used a Bonferroni correction as an adjustment for multiple comparisons in all binomial p-values.

REFERENCES

Adams, M. D., S. E. Celniker, R. A. Holt et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287**:** 2185-2195.

Ashburner, M., C. A. Ball, J. A. Blake et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25**:** 25-29.

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

Boeke, J. D., D. J. Garfinkel, C. A. Styles and G. R. Fink. 1985. Ty elements transpose through an RNA intermediate. Cell 40**:** 491-500.

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

Brosius, J. 1999. Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica 107**:** 209-238.

Castillo-Davis, C. I., F. A. Kondrashov, D. L. Hartl and R. J. Kulathinal. 2004. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. Genome Res 14**:** 802-811.

Celniker, S. E., D. A. Wheeler, B. Kronmiller et al. 2002. Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biol 3**:** RESEARCH0079.

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

Daborn, P. J., J. L. Yen, M. R. Bogwitz et al. 2002. A single p450 allele associated with insecticide resistance in Drosophila. Science 297**:** 2253-2256.

Dimitri, P., N. Junakovic and B. Arca. 2003. Colonization of heterochromatic genes by transposable elements in Drosophila. Mol Biol Evol 20**:** 503-512.

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

Franchini, L. F., E. W. Ganko and J. F. McDonald. 2004. Retrotransposon-Gene Associations Are Wide-Spread among D. melanogaster Populations. Mol Biol Evol**:** msh116.

Ganko, E. W., V. Bhattacharjee, P. Schliekelman and J. F. McDonald. 2003. Evidence for the Contribution of LTR Retrotransposons to C. elegans Gene Evolution. Mol Biol Evol 20**:** 1925-1931.

Ganko, E. W., K. T. Fielman and J. F. McDonald. 2001. Evolutionary history of *Cer* elements and their impact on the *C. elegans* genome. Genome Res 11**:** 2066-2074.

Grover, D., P. P. Majumder, C. B. Rao, S. K. Brahmachari and M. Mukerji. 2003. Nonrandom distribution of alu elements in genes of various functional categories: insight from analysis of human chromosomes 21 and 22. Mol Biol Evol 20**:** 1420-1424.

Harris, M. A., J. Clark, A. Ireland et al. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32 Database issue**:** D258-261.

Hoskins, R. A., C. D. Smith, J. W. Carlson et al. 2002. Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. Genome Biol 3**:** RESEARCH0085-0085.

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

Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 16**:** 418-420.

Kaminker, J. S., C. M. Bergman, B. Kronmiller et al. 2002. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol 3**:** RESEARCH0084-0084.

Leeton, P. R., and D. R. Smyth. 1993. An abundant LINE-like element amplified in the genome of Lilium speciosum. Mol Gen Genet 237**:** 97-104.

Lerat, E., C. Rizzon and C. Biemont. 2003. Sequence Divergence Within Transposable Element Families in the Drosophila melanogaster Genome. Genome Res.**:** 827603.

Li, W. H., Z. Gu, H. Wang and A. Nekrutenko. 2001. Evolutionary analyses of the human genome. Nature 409**:** 847-849.

Makalowski, W. 2000. Genomic scrap yard: how genomes utilize all that junk. Gene 259**:** 61-67.

Maside, X., C. Bartolome and B. Charlesworth. 2002. S-element insertions are associated with the evolution of the Hsp70 genes in Drosophila melanogaster. Curr Biol 12**:** 1686-1691.

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

McDonald, J. F. 1995. Transposable elements: possible catalysts of organismic evolution. Trends Ecol. Evol 10**:** 123-126.

Medstrand, P., J. R. Landry and D. L. Mager. 2001. Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J Biol Chem 276**:** 1896-1903.

Medstrand, P., L. N. van de Lagemaat and D. L. Mager. 2002. Retroelement Distributions in the Human Genome: Variations Associated With Age and Proximity to Genes. Genome Res. 12**:** 1483-1495.

Misra, S., M. A. Crosby, C. J. Mungall et al. 2002. Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. Genome Biol 3**:** RESEARCH0083.

Moriyama, E. N., D. A. Petrov and D. L. Hartl. 1998. Genome size and intron size in Drosophila. Mol Biol Evol 15**:** 770-773.

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

Nigumann, P., K. Redik, K. Matlik and M. Speek. 2002. Many human genes are transcribed from the antisense promoter of l1 retrotransposon. Genomics 79**:** 628-634.

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

Papatsenko, D. A., V. J. Makeev, A. P. Lifanov, M. Regnier, A. G. Nazina and C. Desplan. 2002. Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers. Genome Res 12**:** 470-481.

Petrov, D. A., and D. L. Hartl. 1998. High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. Mol Biol Evol 15**:** 293-302.

Reboul, J., P. Vaglio, N. Tzellas et al. 2001. Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. Nat Genet 27**:** 332-336.

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

Schlenke, T. A., and D. J. Begun. 2004. Strong selective sweep associated with a transposon insertion in Drosophila simulans. Proc Natl Acad Sci U S A 101**:** 1626-1631.

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

Zhang, Z., S. Schwartz, L. Wagner and W. Miller. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol 7**:** 203-214.

Table 4.1 – LTE-gene associations per LTE family

TE Family	LTEs	LTE- gene associations					
		all		internal		proximal	
17.6	12	3	(4)	2	(3)	1	(1)
297	57	24	(19)	15	(12)	9	(7)
412*	31	20	(10)	12	(7)	8	(4)
1731	2	0	(1)	0	(0)	0	(0)
3S18	6	0	(2)	0	(1)	0	(1)
accord	1	1	(0)	1	(0)	0	(0)
aurora-element	3	0	(1)	0	(1)	0	(0)
blastopia	17	10	(6)	7	(4)	3	(2)
blood	22	7	(7)	3	(5)	4	(3)
Burdock	13	8	(4)	7	(3)	1	(2)
Circe	2	0	(1)	0	(0)	0	(0)
copia	30	13	(10)	11	(6)	2	(4)
diver	9	4	(3)	1	(2)	3	(1)
diver2	9	0	(3)	0	(2)	0	(1)
DM88*	32	1	(11)	0	(7)	1	(4)
frogger	1	1	(0)	1	(0)	0	(0)
GATE*	20	0	(7)	0	(4)	0	(2)
gtwin	6	1	(2)	1	(1)	0	(1)
gypsy	2	0	(1)	0	(0)	0	(0)
gypsy2	3	1	(1)	0	(1)	1	(0)
gypsy3	2	0	(1)	0	(0)	0	(0)
gypsy4	2	0	(1)	0	(0)	0	(0)
gypsy5	2	2	(1)	0	(0)	2	(0)
gypsy6	1	0	(0)	0	(0)	0	(0)
HMS-Beagle	13	4	(4)	1	(3)	3	(2)
Idefix	7	2	(2)	0	(1)	2	(1)
invader1*	26	1	(9)	0	(6)	1	(3)
invader2	10	2	(3)	2	(2)	0	(1)
invader3*	16	1	(5)	0	(3)	1	(2)
invader4	9	1	(3)	1	(2)	0	(1)
invader5	6	0	(2)	0	(1)	0	(1)
McClintock	2	1	(1)	0	(0)	1	(0)
mdg1	25	10	(8)	10	(5)	0	(3)
mdg3	16	5	(5)	4	(3)	1	(2)
micropia	5	0	(2)	0	(1)	0	(1)
opus	24	9	(8)	5	(5)	4	(3)
qbert	1	0	(0)	0	(0)	0	(0)
Quasimodo	14	6	(5)	2	(3)	4	(2)
roo	146	58	(49)	42	(31)	16	(18)
rooA	5	2	(2)	1	(1)	1	(1)
rover	6	1	(2)	1	(1)	0	(1)
springer	11	5	(4)	3	(2)	2	(1)
Stalker	12	6	(4)	4	(3)	2	(1)
Stalker2	13	3	(4)	2	(3)	1	(2)
Stalker4	2	0	(1)	0	(0)	0	(0)
Tabor	3	1	(1)	0	(1)	1	(0)
Tirant	20	11	(7)	6	(4)	5	(2)
Transpac	5	3	(2)	1	(1)	2	(1)
	682	228		146		82	

Note: The "all" LTE-gene association column is a combination of LTEs inside predicted gene boundaries ("internal") and LTEs <1000 bp of a gene ("proximal"). Value in parentheses (#) indicates expected value, which is based on the ratio of LTEs in a family / number of all LTEs in the genome, multiplied by the number of associations for a given group. *- observed value significantly different from expectation ($p<.005$)

Table 4.2 – Mean number of genes on *Drosophila* chromosomes per 200 kb region

Chromosome	mean number of genes per 200 kb region	mean for regions with LTEs	mean for regions with LTE-gene associations
2L	21.7 ± 1.9	20 ± 2.1	21.6 ± 2.5
2R	26.1 ± 2.2	24.4 ± 3.6	32 ± 5.1
3L	22.2 ± 2.0	20.4 ± 2.6	23.5 ± 3.1
3R	24.1 ± 1.8	22.5 ± 2.4	23.9 ± 3.2
4	11.6 ± 3.6	13.5 ± 1.7	13.3 ± 2.4
X	20.2 ± 1.7	18.3 ± 1.7	19.7 ± 2.3

Note: Each chromosome was divided into 200 kb regions, and the number of genes in each region was summed. In the second column, the mean number of genes in each region with at least one LTE was calculated. Finally, the mean number of genes in those regions with an LTE-gene association was calculated for the third column. In each case, the mean is followed by the 95% confidence interval.

Table 4.3 – Intron and exon data for *Drosophila* genes with associated LTR retrotransposons

	all genes	genes with proximal LTE association	genes wth internal LTE association
mean intron count per gene	4.4 ± .08	4.9 ± 0.9	9.2 ± 1.1
mean intron size per gene	1085 ± 36	1201 ± 500	4725 ± 535
mean exon count per gene	4.6 ± .07	4.8 ± 1.0	10.0 ± 1.1
mean exon size per gene	487 ± 5	432 ± 54	463 ± 34

Note: In each case, the mean is followed by the 95% confidence interval.

Table 4.4 – "Molecular function" and "cellular component" gene ontology terms for genes associated with LTEs

		LTE- gene associations						
ID	all genes	all assoc		large LTE assoc		small LTE assoc		Ontology description
GO:0003674	**10913**	**305**		**200**		**44**		**molecular_function**
GO:0003754	98	2	(3)	0	(2)	2	(0)	chaperone activity
GO:0003774	79	0	(2)	0	(1)	0	(0)	motor activity
GO:0003793	64	2	(2)	2	(1)	0	(0)	defense/immunity protein activity
GO:0003824	4504	122	(126)	101	(83)	6	(18)	catalytic activity
GO:0004871	845	62*	(24)	21	(15)	19*	(3)	signal transducer activity*
GO:0005194	60	2	(2)	2	(1)	0	(0)	cell adhesion molecule activity
GO:0005198	339	5	(9)	5	(6)	0	(1)	structural molecule activity
GO:0005215	1180	24	(33)	15	(22)	1	(5)	transporter activity
GO:0005488	2429	59	(68)	33	(45)	12	(10)	binding
GO:0005554	9	0	(0)	0	(0)	0	(0)	molecular_function unknown
GO:0008369	260	2	(7)	0	(5)	2	(1)	obsolete
GO:0008580	1	0	(0)	0	(0)	0	(0)	cytoskeletal regulator activity
GO:0008638	6	0	(0)	0	(0)	0	(0)	protein tagging activity
GO:0016209	10	0	(0)	0	(0)	0	(0)	antioxidant activity
GO:0016329	17	1	(0)	1	(0)	0	(0)	apoptosis regulator activity
GO:0030188	1	0	(0)	0	(0)	0	(0)	chaperone regulator activity
GO:0030234	268	3	(7)	2	(5)	0	(1)	enzyme regulator activity
GO:0030528	668	20	(19)	17	(12)	2	(3)	transcription regulator activity
GO:0045182	70	1	(2)	1	(1)	0	(0)	translation regulator activity
GO:0045735	5	0	(0)	0	(0)	0	(0)	nutrient reservoir activity
		χ2 p=	8.3E-11		0.150		9.5E-19	
GO:0005575	**6331**	**142**		**116**		**17**		**cellular_component**
GO:0005576	205	8	(5)	8	(4)	0	(1)	extracellular
GO:0005623	5983	132	(134)	106	(110)	17	(16)	cell
GO:0005941	76	1	(2)	1	(1)	0	(0)	unlocalized
GO:0008370	52	0	(1)	0	(1)	0	(0)	obsolete
GO:0008372	15	1	(0)	1	(0)	0	(0)	cellular_component unknown
		χ2 p=	0.504		0.273		0.900	

Note: Shown are counts of gene ontology descriptor terms for all genes associated with an LTE element ("all assoc"), genes associated with a large LTE element (>90% of consensus size, "large LTE assoc"), or genes associated with a small LTE element (<20%

of consensus size, “small LTE assoc"). Value in parentheses (#) indicates expected value based on the ratio of descriptor GO terms for all *Drosophila* genes ("all genes") in the root molecular_function and cellular_component ontologies. *- observed value significantly different from binomial expectation (p<.1.0e-5)

Table 4.5 – “Biological process” gene ontology terms for genes associated with LTEs

		LTE- gene associations						
ID	all genes	all assoc		large LTE assoc		small LTE assoc		Ontology description
GO:0008150	**18299**	**650**		**546**		**84**		**biological_process**
GO:0000004	7	1	(0)	1	(0)	0	(0)	biological_process unknown
GO:0007275	4663	225*	(166)	200*	(139)	20	(21)	development*
GO:0007582	9257	255*	(329)	204*	(276)	44	(42)	physiological processes*
GO:0007610	253	32*	(9)	32*	(8)	0	(1)	behavior*
GO:0008371	31	1	(1)	1	(1)	0	(0)	obsolete
GO:0009987	4087	136	(145)	108	(122)	20	(19)	cellular process
GO:0016032	1	0	(0)	0	(0)	0	(0)	viral life cycle
		χ2 p=	4.4E-20		8E-25		0.75409	

Note: Shown are counts of gene ontology descriptor terms for all genes associated with an LTE element ("all assoc"), genes associated with a large LTE element (>90% of consensus size, "large LTE assoc"), or genes associated with a small LTE element (<20% of consensus size, “small LTE assoc"). Value in parentheses (#) indicates expected value based on the ratio of descriptor GO terms for all *Drosophila* genes ("all genes") in the root molecular_function and cellular_component ontologies. *- observed value significantly different from binomial expectation (p<.1.0e-5)

Table 4.6 – Distribution of "development", "physiological process" and "behavior" gene ontology terms for genes assoctiated with LTEs

ID	all genes	all assoc		Ontology description
GO:0007275	**4663**	**102**		**development**
GO:0000003	383	15	(18)	reproduction
GO:0002164	327	25	(16)	larval development
GO:0007320	5	0	(0)	insemination
GO:0007349	10	0	(0)	cellularization
GO:0007389	673	8*	(32)	pattern specification
GO:0007530	59	0	(3)	sex determination
GO:0007548	10	2	(0)	sex differentiation
GO:0007568	9	0	(0)	aging
GO:0009292	3	0	(0)	genetic transfer
GO:0009653	1904	115*	(92)	morphogenesis
GO:0009790	442	15	(21)	embryonic development
GO:0009791	217	15	(10)	post-embryonic development
GO:0019827	3	0	(0)	stem cell maintenance
GO:0030154	437	22	(21)	cell differentiation
GO:0040007	37	0	(2)	growth
GO:0048066	125	8	(6)	pigmentation
GO:0040029	17	0	(1)	regulation of gene expression, epigenetic
		$\chi 2$ p=	2E-05	
GO:0007582	**9257**	**92**		**physiological processes**
GO:0006950	285	15	(8)	response to stress
GO:0007586	4	0	(0)	digestion
GO:0007588	1	0	(0)	excretion
GO:0008015	4	0	(0)	circulation
GO:0008151	2424	69	(67)	cell growth and/or maintenance
GO:0008152	4747	94*	(131)	metabolism
GO:0009605	1335	69*	(37)	response to external stimulus
GO:0009719	78	0	(2)	response to endogenous stimulus
GO:0016265	111	1	(3)	death
GO:0030431	7	0	(0)	sleep
GO:0042303	26	2	(1)	molting cycle
GO:0042592	21	0	(1)	homeostasis
GO:0046903	210	5	(6)	secretion
		$\chi 2$ p=	4E-08	
GO:0007610	**253**	**32**		**behavior**
GO:0007611	41	10	(5)	learning and/or memory
GO:0007622	49	2	(6)	rhythmic behavior
GO:0007625	2	1	(0)	grooming behavior
GO:0007626	34	3	(4)	locomotory behavior
GO:0007631	4	2	(1)	feeding behavior
GO:0007635	2	0	(0)	chemosensory behavior
GO:0007638	3	0	(0)	mechanosensory behavior
GO:0019098	66	4	(8)	reproductive behavior
GO:0030534	27	4	(3)	adult behavior
GO:0030537	14	3	(2)	larval behavior
GO:0040040	1	0	(0)	thermosensory behavior
		$\chi 2$ p=	0.068	

Notes: Shown are counts of subordinate descriptors from three biological_process terms for genes associated with an LTE element. Parentheses (#) indicates expected value based on the ratio of descriptor GO terms from all Drosophila genes ("all genes") in the biological_process ontology. *- observed value significantly different from binomial expectation ($p<0.05$)

Figure 4.1: Size distribution of LTR retrotransposons associated with genes. Full-length consensus sizes for individual families were determined from Flybase (Kaminker et al. 2002) or RepBase (Jurka 2000), and the size of each LTE element was compared to the size of the consensus sequence for the appropriate LTE family to calculate a "percent consensus size." Near full-length LTEs are ≥90% of the consensus size, medium sized LTEs are 21% - 89% of consensus size, and small LTEs are ≤20% of consensus size. Expected values were calculated based on the distribution of all LTEs in the genome. Stars indicate that observed values are significantly different than the expected value ($p<0.05$).

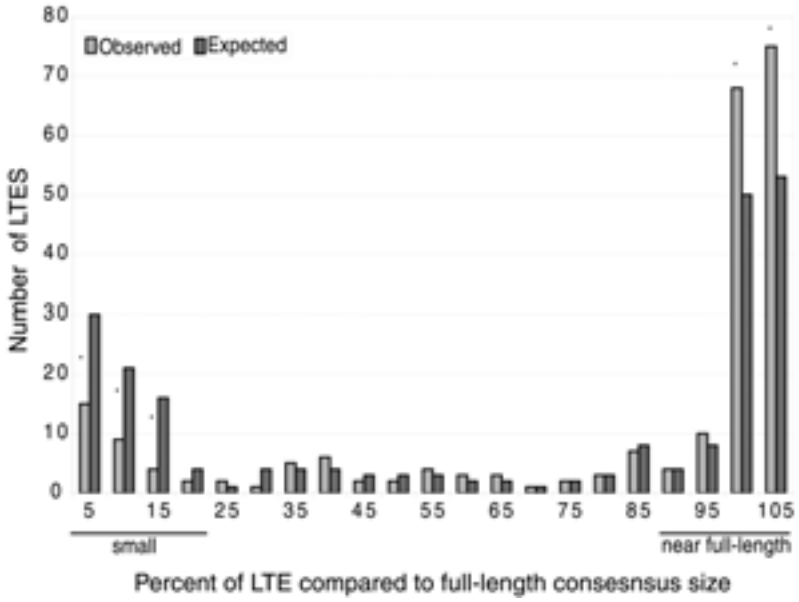

Observed
Expected
Number of LTES
80
70
60
50
40
30
20
10
0
5
15
25
35
45
55
65
75
85
95
105
small
near full-length
Percent of LTE compared to full-length consesnsus size

Figure 4.2: Size composition of LTR retrotransposons near *Drosophila* genes.

The mean size of LTEs was computed for all LTEs, for LTEs proximal to a gene (1-1000 bp), for LTEs inside a gene, and for LTEs in or proximal to conserved homologenes. Error range indicates 95% confidence interval. Each bar was shaded based on the number of LTEs from three size groups: near full-length (LTEs ≥90% of the consensus size), medium sized (21% - 89% of consensus size), and small sized (LTEs ≤20% of consensus size).

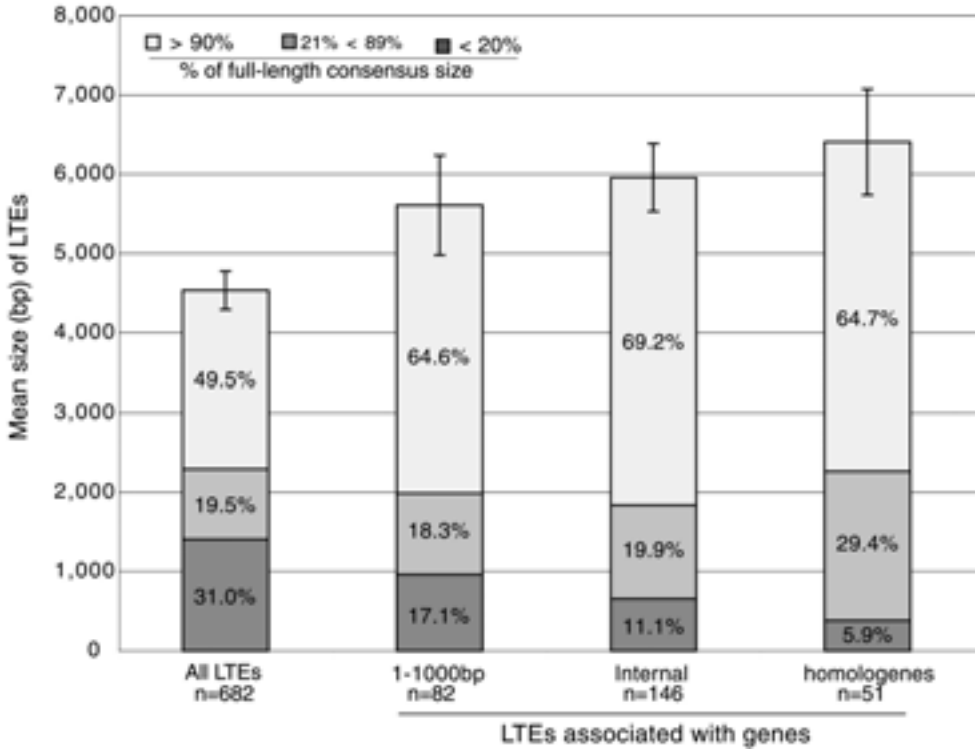
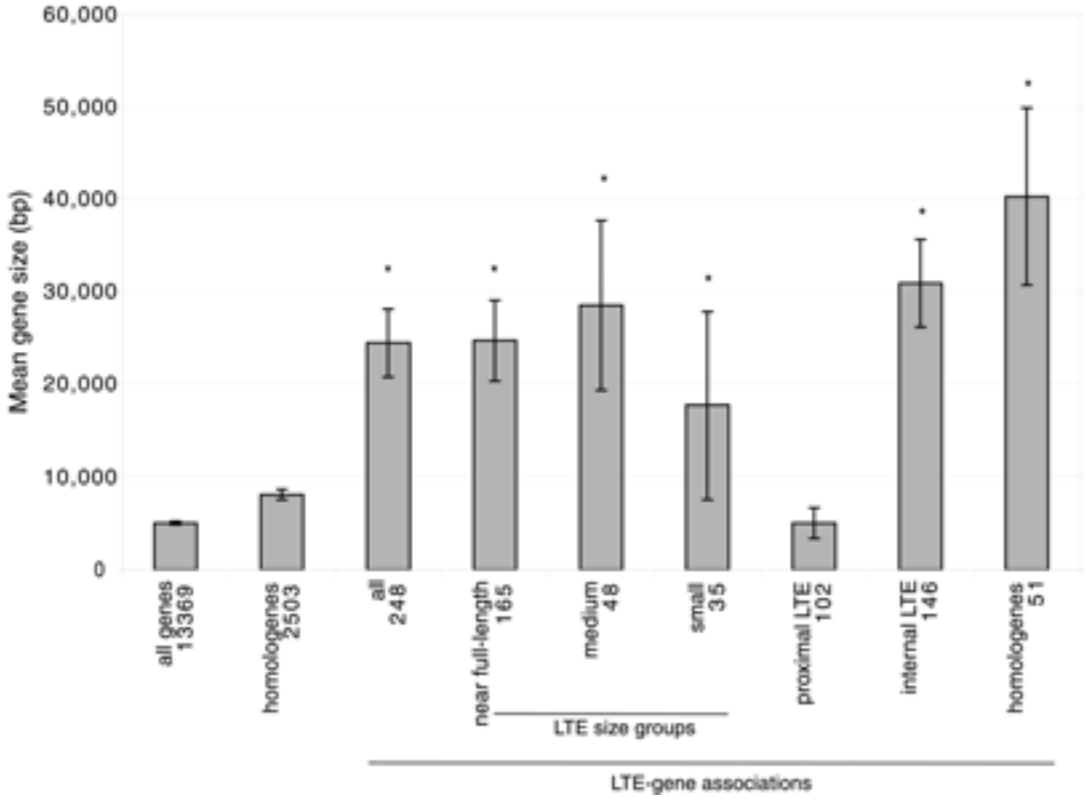

8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
Mean size (bp) of LTEs
> 90%
21% < 89%
< 20%
% of full-length consensus size
49.5%
19.5%
31.0%
64.6%
18.3%
17.1%
69.2%
19.9%
11.1%
64.7%
29.4%
5.9%
All LTEs
n=682
1-1000bp
n=82
Internal
n=146
homologenes
n=51
LTEs associated with genes

Figure 4.3: Mean size of *Drosophila* genes in or proximal to LTR retrotransposons. Each bar represents the mean size for a set of genes. The three LTE size groups are based on comparisons to consensus size: near full-length (LTEs ≥90% of the consensus size), medium sized (21% - 89% of consensus size), and small (LTEs ≤20% of consensus size). Proximal associations are LTEs within 1-1000 bp of gene, and internal associations are LTEs inside a gene. Error range indicates 95% confidence interval. Stars indicate a significantly different gene size for a particular group of genes associated with an LTE compared to the mean size of all genes ($p<0.05$).


60,000
50,000
40,000
30,000
20,000
10,000
0
Mean gene size (bp)
all genes 13369
homologenes 2503
all 248
near full-length 165
medium 48
small 35
proximal LTE 102
internal LTE 146
homologenes 51
LTE size groups
LTE-gene associations

CHAPTER 5

CONCLUSION

Transposable elements have a broad range of effects in the genome, ranging from mutation and disease, to increased rates of recombination, to becoming functional components of genes. LTR retrotransposons are class I transposons with regulatory features such as promoters, enhancers, and termination signals, properties that lend them the potential for functional adaptation by a host organism. To better understand the potential for functional adaptation, I have analyzed the genomes of two sequenced multicellular animals, *C. elegans* and *D. melanogaster*, for the presence of LTR retrotransposons in and proximal to genes.

New *Cer* (*C. elegans* retrotransposon) element families and subfamilies were identified in the course of the *C. elegans* analysis. Identification of shared primer binding sites (PBS) between related families provided further support for the evolutionary classification of *Cer* elements. Fragmented elements and solo LTRs make up the majority of *Cer* elements. In fact, several families consist entirely of these putatively older, "dead" elements. The high percentage of *Cer* fragments and solo LTRs is likely due to rates of deletion (Kent and Zahler 2000) and especially recombination (Coghlan and Wolfe 2002) in the *C. elegans* genome.

Two-thirds of *Cer* elements are in or near genes. Most of these retrotransposon associated genes are on the chromosome arms, a region of the *C. elegans* genome with a higher concentration of pseudogenes, duplicated genes and chimeric genes (Consortium

1998). Many genes on the chromosome arms are predicted to be nematode-specific genes with a recent evolutionary history. *Cer* elements that insert on the chromosome arms are less likely to disrupt genes with a critical function since those genes tend to be located in the central region of the chromosome. Elements on the chromosome arms may also promote recombination or provide regulatory and coding components to evolving genes. It is possible that the chromosomal arms of *C. elegans* are an "evolutionary laboratory" where new genes are created and tested by natural selection.

In Drosophila euchromatin one-third of LTR retrotransposons are located inside genes or within the predicted regulatory boundaries of genes. Many of these element-associated genes have functions related to interactions outside the cell, while genes with metabolic and cell differentiation functions have fewer neighboring transposons. Genes with external functions may be genetically malleable in order to evolve in response to new external challenges, and this malleability may allow transposons near these genes to be tolerated or even functionally adapted. Metabolic and cell differentiation functions are less malleable and thus more prone to insertional disruption, so transposons near these genes are more likely to face strong negative selection.

The skewed size distribution of retrotransposons associating with genes is evidence of selection acting on retrotransposons in Drosophila. Most of the retrotransposons associated with genes are large and evidence indicates that many of these large retrotransposons are relatively recent insertions (Bowen and McDonald 2001; Kaminker et al. 2002). A recently published study using data from Chapter 4 has shown that large transposons are confined to the sequenced *D. melanogaster* strain, while some small transposons were found in populations around the world (Franchini, Ganko and

McDonald 2004). Essentially, most large transposons are recent insertion events and have not yet been affected by selective evolution. It is likely that many of these elements are non-adaptive and will face future degradation and eventual elimination. Older elements that were disruptive to genes have been deleted over time, with the remaining small elements unlikely to have a major disruptive effect. Both large and small elements with adaptive functionality are expected to face reduction to a functional core over evolutionary time.

Transposons in *C. elegans* and *D. melanogaster* are in relatively low abundance (<5% of the genome) compared to the human genome (>40% of the genome). Accordingly, humans have a greater number of transposons providing potential genetic plasticity through interactions with host genes. The tradeoff to increased plasticity is a much larger genome in humans (>3 Bbp) compared to *C. elegans* (100 Mbp) and *D. melanogaster* (180Mbp). For example, there are a number of human-specific transposon insertions (Medstrand and Mager 1998) and the human genome is estimated to have increased 20% in size through the primate lineage in large part due to transposon insertion (Liu et al. 2003). In *C. elegans*, deletion, high recombination rates (Coghlan and Wolfe 2002; Holt et al. 2002) and few full-length elements make it difficult for LTR retrotransposons to achieve high copy number. A general lack of solo LTRs in Drosophila (Kaminker et al. 2002) indicates recombination is less important in transposon control than in the *C. elegans* genome. Deletion is probably the major transposon degradation force in Drosophila, as only the most recently inserted elements lack deletions. Indeed, the *D. melanogaster* genome appears to have shrunk over time compared to other insects indicating deletion on a large scale level (Holt et al. 2002).

Though worms, flies and humans have drastically different levels of transposon content and respond differently to transposon insertion, all three do have confirmed cases of transposons contributing adaptive functionality.

The ultimate adaptiveness of a particular transposon on a gene is difficult to discern. Identifying and analyzing potential sites is a first step, and this dissertation has presented a number of targets in two sequenced model organisms. Further molecular and population studies on identified targets will lead to a better understanding of LTR retrotransposon contribution to gene evolution in *C. elegans* and *D. melanogaster*.

REFERENCES

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

Coghlan, A., and K. H. Wolfe. 2002. Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. Genome Res 12**:** 857-867.

Consortium, *C. e. S.* 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. The *C. elegans* Sequencing Consortium. Science 282**:** 2012-2018.

Franchini, L. F., E. W. Ganko and J. F. McDonald. 2004. Retrotransposon-Gene Associations Are Wide-Spread among *D. melanogaster* Populations. Mol Biol Evol**:** msh116.

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

Kaminker, J. S., C. M. Bergman, B. Kronmiller et al. 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol 3**:** RESEARCH0084-0084.

Kent, W. J., and A. M. Zahler. 2000. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae-C. elegans* genomic alignment. Genome Res 10**:** 1115-1125.

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

Medstrand, P., and D. L. Mager. 1998. Human-specific integrations of the HERV-K endogenous retrovirus family. J Virol 72**:** 9782-9787.