Fall 1996

Variation in transposable element sequence and activity in the nematode Caenorhabditis elegans

Jeremy David Glasner
University of New Hampshire, Durham

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

VARIATION IN TRANSPOSABLE ELEMENT SEQUENCE AND ACTIVITY IN THE NEMATODE *CAENORHABDITIS ELEGANS*

BY

JEREMY D. GLASNER
B.S., BIOLOGY, PENNSYLVANIA STATE UNIVERSITY, 1991

DISSERTATION

Submitted to the University of New Hampshire
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

in

Genetics

September, 1996

This dissertation has been examined and approved.

Dissertation Co-Director, John J. Collins
Associate Professor of Biochemistry and
Molecular Biology and Graduate Program
in Genetics

Dissertation Co-Director, Thomas D. Kocher
Associate Professor of Zoology and Graduate
Program in Genetics

Anita S. Klein
Associate Professor of Biochemistry and
Molecular Biology and Chair of Graduate
Program in Genetics

Clyde L. Denis
Professor of Biochemistry and Molecular
Biology and Graduate Program in Genetics

Marianne K. Litvaitis
Assistant Professor of Zoology

7/18/96
Date

ACKNOWLEDGEMENTS

I would like to thank my dissertation advisors, John Collins and Tom Kocher, for their participation, interest and support. I am also grateful to the rest of my committee, Clyde Denis, Anita Klein, and Marian Litvitis, for their contributions to my completion of this degree and education as a whole. Many other faculty of the Genetics Program, Department of Biochemistry and Department of Zoology have been influential. I greatly appreciate the encouragement my numerous labmates, friends and collegues. I would also like to thank the Graduate School for support throughout the tenure of my stay at the University of New Hampshire. Special thanks to my extended family for unwaivering confidence and to Nicole, for being Nicole.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iv
LIST OF TABLES viii
LIST OF FIGURES x
ABSTRACT xii

CHAPTER I

INTRODUCTION TO TRANSPOSABLE ELEMENTS AND GENOME EVOLUTION 1
General Introduction 1
Mutation underlies genetic variation 2
Transposons as a source of genetic variation 3
Types of transposons 3
Transposons can generate major chromosomal rearrangements 6
Changes in gene sequences 6
Transposons affect splicing of RNA transcripts 8
Transposons affect gene regulation 11
Regulation of transposable element activity 13
Transposons and evolution 16
Levels of Selection 17
The evolution of transposon sequences 19
The evolution of genomes containing transposons 21
Methods for understanding the madness 22
Transposons in *C. elegans* 25
C. elegans as a model 25
Thesis organization 27

CHAPTER II

TRANSPOSONS IN THE *C. ELEGANS* GENOME: VARIATION WITHIN AND BETWEEN ELEMENT FAMILIES 28
Introduction 28
Discovery of transposons in *C. elegans* 28
Distribution of element insertion sites across the genome 31
Transposon-like sequences in the *C. elegans* genome sequence 32
Methods 34
Results and Discussion 37
Genomic Distribution 37
Analysis of Tc1 and Tc1-like sequences in the *C. elegans* genome . . . 42
Analysis of Tc2 and Tc2-like sequences in the *C. elegans* genome . . . 50
Analysis of Tc3 and Tc3-like sequences in the *C. elegans* genome . . . 55
Analysis of Tc4 and Tc4-like sequences in the *C. elegans* genome . . . 61
Analysis of Tc5 and Tc5-like sequences in the *C. elegans* genome . . . 65
Analysis of Tc6 and Tc6-like sequences in the *C. elegans* genome . . . 72

Conclusions 79

CHAPTER III

ATTEMPTS TO CHARACTERIZE THE PHENOTYPIC CONSEQUENCES OF TRANSPOSABLE ELEMENT INSERTION 83
Summary 83
Introduction 83
Genetic methods may underestimate the level of transposon activity and the range of phenotypic variation elements can generate. . . 84
Sib-selection/PCR can isolate insertions without regard to phenotype . . 85
Muscle genes are good targets 86
Tc1 is active in the germline of *mut-2* animals 87
Methods 87
Sib-selection PCR with Southern blotting to isolate Tc1 insertions in *unc-54* 88
Sib-selection PCR with nested PCR to isolate Tc1 insertions in *unc-54* and *unc-22* 89
Choice of a strain for sib-selection 91
Results 91
PCR and Southern Blotting to detect insertions 91
Sib-selection with nested PCR to detect insertion events 93
unc-22 is a difficult target for detecting new insertions by PCR 94
unc-54::Tc1 insertions are detected by PCR but are difficult to isolate by sib-selection 95

CHAPTER IV

HIGH FREQUENCY SOMATIC INSERTION OF TC1 IN *C. ELEGANS* 101
Introduction 102
Materials and Methods 105
C elegans strains and maintenance 105
DNA extraction and PCR amplification 106
Genomic Southern blots 108
Detection of insertions in parents and their offspring 108
Sequencing of PCR products 109
Construction of strains that contain somatic Tc1 activity and the *glp-4(bn2)* allele 109
Laser ablation of TW332 larvae 110
Results 110
Tc1 insertion into the *unc-54* gene occurs frequently in TW332 110
The high frequency of Tc1 insertion into *unc-54* occurs in most wild-type genetic backgrounds 117
Tc1 insertions arise during culture of TW332 and EM1002, and are not inherited 117
Tc1 insertions into *unc-54* are detected in adult worms lacking a germline 121
The sequence of the *unc-54* hotspot varies between strains 126
Sequences of insertion sites 127

Tc1 inserts frequently into another region of *unc-54* 129
Another *C. elegans* gene, src-1, contains hotspots for somatic insertion of Tc1 129
Discussion and Conclusions 130
Tc1 inserts at high frequency in somatic cells 130
Regulation of somatic Tc1 activity 131
What makes a hotspot hot? 135
Somatic transposition and reverse genetics 137
Evolutionary significance of somatic transposition 139

LIST OF REFERENCES 142

APPENDICES 153
APPENDIX A: Alignment of Tc1 and seven cosmid sequences identified as high scoring blast hits to Tc1 153
APPENDIX B: Alignment of seven additional cosmid sequences identified as high scoring blast hits to Tc1 159
APPENDIX C: Alignment of a modified Tc2 sequence and ten cosmid sequences identified as high scoring blast hits to Tc2 162
APPENDIX D: Alignment of Tc3 and ten cosmid sequences identified as high scoring blast hits to Tc3 165
APPENDIX E: Alignment of Tc4 and six cosmid sequences identified as high scoring blast hits to Tc4 177
APPENDIX F: Alignment of Tc5 and the cosmid T13c2 sequence identified as high scoring blast hit to Tc5 182
APPENDIX G: Alignment of four cosmid sequences identified as high scoring blast hits to Tc5 190
APPENDIX H: Alignment of five short cosmid sequences identified as high scoring blast hits to Tc5 193
APPENDIX I: Alignment of Tc6 and nine cosmid sequences identified as high scoring blast hits to Tc6 195

LIST OF TABLES

Table 2.1: Genomic location of transposon-like sequences in the *C. elegans* genome. 38

Table 2.2: Comparison of the length of IR, variation among the two IR of each element, and total length for Tc1 and cosmid sequences of high-scoring BLAST hits. 43

Table 2.3: Pairwise distances between Tc1 and cosmid sequences for positions 11-1621 of the APPENDIX A alignment. Absolute distance are shown in the lower diagonal. Mean distances (adjusted for missing data) are shown in the upper diagonal. 44

Table 2.4: Variable sites from an alignment of predicted transposases for Tc1 and seven Tc1-like elements. 45

Table 2.5: Pairwise distances between Tc1-like cosmid sequences for positions 11-936 of the APPENDIX B alignment. 48

Table 2.6: Comparison of the length of IR, variation among the two IR of each element, and total length for Tc2 and Tc2-like cosmid sequences. 50

Table 2.7: Lists the position of gaps found among Tc2 related sequences in the alignment shown in Appendix C. 51

Table 2.8: Pairwise distances between Tc2-like cosmid sequences for positions 11-499 of the APPENDIX C alignment. 54

Table 2.9: Comparison of the length of IR, variation among the two IR of each element, and total length for Tc3 and Tc3-like cosmid sequences. 56

Table 2.10: Pairwise distances between Tc3 and the four cosmid sequences with greatest similarity to Tc3 from the APPENDIX D alignment.. 59

Table 2.11: Pairwise distances between three sequences from the APPENDIX D alignment that are shorter than Tc3 and encode a predicted protein that is similar to the Tc3 transposase. 59

Table 2.12: Variable sites from an alignment of predicted transposases for Tc3 and three Tc3-like elements. 60

Table 2.13: Pairwise distances between Tc3 transposase and the predicted amino acid sequence from four shorter cosmid sequences that also encode a signifcant ORF. 60

Table 2.14: Comparison of the length of IR, variation among the two IR of each element, and total length for Tc4 and Tc4-like cosmid sequences. 62

Table 2.15 Pairwise distances between Tc4 and the six Tc4-like cosmid sequences from the APPENDIX E alignment. 63

Table 2.16: Comparison of the length of IR, variation among the two IR of each element, and total length for Tc5 and Tc5-like cosmid sequences. 65

Table 2.17: Describes the position of insertions and deletions among Tc5 related elements from the alignment in Appendix G. 68

Table 2.18: Pairwise distances between Tc5-like cosmid sequences for positions 17-1653 of the APPENDIX G alignment. 68

Table 2.19: Describes the position of insertions and deletions among Tc5 related elements from the alignment in Appendix H. 70

Table 2.20: Pairwise distances between five short Tc5-like cosmid sequences for positions 12-717 of the APPENDIX H alignment. 72

Table 2.21: Comparison of the length of IR, variation among the two IR of each element, and total length for Tc6.1 and Tc6-like cosmid sequences. 73

Table 2.22: Describes insertions and deletions among Tc6 related elements from the alignment contained in Appendix I. 75

Table 2.23: Pairwise distances between Tc6.1 and Tc6-like cosmid sequences for positions 12-1627 of the APPENDIX I alignment. 76

Table 4.1 Sequences of PCR primers used to detect Tc1 insertions. 106

Table 4.2 Summary of somatic insertion frequencies in different strains and life stages. 113

Table 4.3 Summary of insertion frequencies in parental worms and their larval and adult offspring. 120

Table 4.4 Summary of somatic insertion frequencies in *glp-4(bn2)* strains. 122

Table 4.5 Summary of somatic insertion frequencies in TW332 animals with germlines ablated and without ablation. 123

LIST OF FIGURES

Figure 1.1 Basic structure of two major classes of transposable elements 5

Figure 1.2 Splicing patterns observed for the wild-type *unc-22* gene and two *unc-22*::Tc3 alleles 44

Figure 2.1 Diagram showing the position of transposon-like sequences on the major contig for each chromosome 41

Figure 2.2 Parsimony bootstrap consensus tree of 8 Tc1 elements 47

Figure 2.3 Parsimony bootstrap consensus tree of foldback Tc1-like elements 49

Figure 2.4 Parsimony bootstrap consensus tree of Tc2-like elements 52

Figure 2.5 Parsimony bootstrap consensus tree of Tc2del and related elements 53

Figure 2.6 Parsimony bootstrap consensus tree of Tc3 and related elements 58

Figure 2.7 Parsimony bootstrap consensus tree of Tc4 and related elements 64

Figure 2.8 Parsimony bootstrap consensus tree of Tc5del and related elements 66

Figure 2.9 Parsimony bootstrap consensus tree of Tc5-like foldback elements 69

Figure 2.10 Parsimony bootstrap consensus tree of short Tc5-like elements 71

Figure 2.11 Parsimony bootstrap consensus tree of Tc6 and related elements 77

Figure 2.12 Parsimony bootstrap consensus tree of complete Tc6-like elements 78

Figure 3.1 Location of *unc-54* and Tc1 primers used in PCR experiments 90

Figure 3.2 Flow chart for detection and sib-selection of an insertion detected with JC68 and JC69. 97

Figure 4.1 Location of PCR primers in *unc-54* gene and Tc1 transposon. 107

Figure 4.2 Agarose gel showing typical PCR products amplified from single animals. 111

Figure 4.3 Genomic Southern Blot of *BamH1* digested DNAs from N2, TR1299 and TW332 worms and probed with punk-54, a cloned copy of *unc-54*. . . . 116

Figure 4.4 PCR products from single animals amplified with nested primers JC58 and JC67. 119

Figure 4.5 PCR products amplified from single adult hermaphrodites. 124

Figure 4.6 The diagram shows the exon3/intron3 boundary in the *unc-54* gene. . . 125

ABSTRACT

VARIATION IN TRANSPOSABLE ELEMENT SEQUENCE AND ACTIVITY IN THE NEMATODE *CAENORHABDITIS ELEGANS*

by

Jeremy D. Glasner
University of New Hampshire, September, 1996

Eukaryotic genomes are replete with transposable elements. The nematode *C. elegans* will be the first multicellular organism to have its genome completely sequenced. This sequence will allow identification of all the transposon and transposon-related sequences from a single genome. In anticipation of the complete genome sequence I have initiated a series of analyses of sequences from the *C. elegans* genome database that share significant similarity to known families of transposons. Several members of known transposon families were observed along with a plethora of sequences related to these known transposons. Cladistic analyses were used to describe the relationships among transposons and transposon families. These analyses suggest that transposons in *C. elegans* may be found in both autonomous and nonautonomous forms. The differences between related element families lies mostly in the length of the inverted repeats and the presence of open reading frames. Differences between sequences within an element family suggest several mechanisms for generating length variation in inverted repeats.

Characterization of the consequences of Tc1 insertion requires a means of detecting insertions. I describe reverse genetic methodology for identifying new transposon insertions. To study the regulation of transposon activity I focused on the tissue-specific and developmental regulation of Tc1. I identified sites that are frequent targets for Tc1 insertion. In the most dramatic example, insertion of Tc1 was detected at the same site in the *unc-54* gene in nearly every animal screened. This site was previously shown to be a

"hotspot" for germ-line insertion, although at a frequency several orders of magnitude less than the levels now detected. I believe these insertions are somatic events because they increase in frequency during development but are not transmitted to progeny based on both genetic and molecular evidence and because I detect them in animals lacking a germline. Additional sites in *unc-54* and *src-1*, another *C. elegans* gene, were identified as frequent targets for insertion of Tc1; however, none are hit as frequently as the *unc-54* "hotspot". Somatic insertion of Tc1 depends on genetic background and may be suppressed early in development.

CHAPTER I

INTRODUCTION TO TRANSPOSABLE ELEMENTS AND GENOME EVOLUTION

General Introduction:

There is an amazing abundance and diversity of life in our environment. As fellow creatures on this planet we have a natural interest and wariness of the life that surrounds us. Humans often consider themselves unique among animals because of their ability to reason and contemplate their own existence. Our curiosity has lead to great strides in understanding the origin of life and the complexities of its workings. The single greatest leap in our understanding of life is the realization that all living things share a common origin and that the process responsible for the amazing diversity of life is evolution.

Thousands of independent pieces of evidence lead to the conclusion that evolution is a biological fact. All life comes from life, and ultimately, all species come from other species through a process of descent with modification. Given that evolution happens, one goal of biological research is to understand how it occurs. Perhaps the greatest contribution to this understanding comes from the field of genetics.

As evolutionary biologists we want to know how organisms develop and reproduce giving rise to individuals of the same species. In addition, we want to know how diversity arises among individuals of the same species and how this relates to the diversity observed between species. A mechanistic explanation for inheritance and diversity comes from an understanding of genetics and molecular biology. Inheritance implies the presence of parental characteristics in the next generation. Research conducted by numerous scientists over the last few decades has demonstrated that DNA is the vehicle that carries the information necessary for development and reproduction. The knowledge that DNA

provides the basis for inheritance lead to the realization that many of the differences between individuals, as well as differences between species, arise from changes in their DNA sequences. Evolution occurs because of changes in the frequencies of different DNA sequence variants within a population. In some cases changes in the frequency of a variant will be driven by a selective difference between individuals with different genotypes, and in other cases changes will occur through chance fluctuations in frequency (genetic drift).

Fundamental to the goal of understanding the process of evolution is characterization of the mechanisms that alter DNA sequences. Ultimately, all heritable variation must arise from changes occurring at the DNA level. So, to understand the nature of genetic variation it is necessary to examine the process of mutation.

Mutation underlies genetic variation:

Mutations come in many varieties and can be observed at many different levels. Some mutations are "silent", they affect the DNA sequence of an individual but result in no observable change in the individual. Other mutations have dramatic consequences for an individual and in the most extreme cases are lethal. Occasionally a mutation may arise that provides an individual with an advantage in survival or reproduction. Some mutations may be silent under one set of conditions but deleterious or beneficial under another. The consequences of mutation are complicated and variable. The causes of mutation are complicated and variable as well, but are understood to a greater extent than their consequences.

A multitude of mutational mechanisms introduce genetic variation. Mutations arise as changes in DNA sequences. There are several basic types of mutation. Single base substitutions alter one nucleotide position at a time. Insertions result in the addition of one or more bases into an existing sequence. Deletions lead to loss of one or more bases from a sequence. Chromosomal rearrangements affect large pieces of DNA. Some chemicals,

know as mutagens, are know to induce particular types of mutation. For example, ethane methyl sulfonate (EMS) is known to lead to a increase in the frequency of particular single base substitutions in DNA sequences. Many mutagens result in mutation only after DNA replication occurs. The initial lesion generated by the mutagen does not lead to a change in the DNA sequence until it is replicated. Other mutagenic agents found in the environment, such as X-rays, increase the frequency of chromosomal rearrangement. Mutagens can be thought of as external factors that lead to mutation. Many mutations arise from processes that are a normal part of cellular activity. DNA replication itself can lead to mutation, as when an incorrect base is placed in a replicating DNA molecule. Errors in repair or recombination of DNA can also lead to mutation. A particularly intriguing mutational pathway results from the action of endogenous transposable elements.

Transposable elements, or transposons, are ubiquitous components of prokaryotic and eukaryotic genomes. Transposons are DNA sequences found in multiple copies within a cell. The cardinal feature of transposable elements is their ability to move within their host genome (reviewed in Berg and Howe, 1989; Lambert et al., 1989) . Transposons can insert into previously unoccupied DNA sequences and excise from sites that they occupy. Insertion and excision of transposons can lead to almost any type of mutation including single base substitutions, insertions and deletions, and chromosomal rearrangements. Since transposons are ubiquitous and generate many types of mutation, they are likely to play a unique and important role in molecular evolution.

Transposons as a source of genetic variation:

Types of transposons:

Eukaryotic transposons are often divided into two basic classes by their mechanism of transposition (reviewed in Finnegan, 1989). Class I elements, often referred to as retrotransposons, transpose through an RNA intermediate. Retrotransposon sequences are

transcribed into RNA, reverse transcribed into cDNA and inserted into a new genomic location. Class II elements are thought to transpose directly from DNA to DNA without an RNA intermediate.

In addition to the differences in mechanisms of transposition, the elements are structurally distinct as illustrated in Figure 1.1. In general, class I elements are similar to endogenous retroviruses (reviewed in Berg and Howe, 1989; Lambert et al, 1989). These retrotransposons include Ty elements in *Saccharomyces cerevisiae*, copia-like elements in *Drosophila melanogaster*, and myriad elements in organisms as diverse as *Zea mays* and humans. Most of these elements have long terminal direct repeats (LTRs) flanking sequences containing long open reading frames, one of which encodes a reverse transcriptase-like product. Other class I elements lack direct repeats at their termini but also contain gag and reverse transcriptase like coding regions. Still others, known as SINEs, lack LTRs and coding sequences but require reverse transcriptase activity to move. Class II elements resemble insertion sequences in prokaryotes. Class II transposons include P elements in *Drosophila*, *Ac/Ds* elements in maize, Tc1 elements in *Caenorhabditis elegans*, and multiple sequences from all eukayotic genomes where they have been sought (reviewed in Berg and Howe, 1989). Class II transposons are generally identified as sequence elements with inverted repeats at their termini. Some elements, referred to as foldback elements, are almost entirely inverted repeat sequences. Other class II transposons contain one or more open reading frames that encode products called transposases that are thought to be involved in the transposition process. Some class II elements, referred to as nonautonomous or defective transposons and are related to elements in the genome that encode transposases.

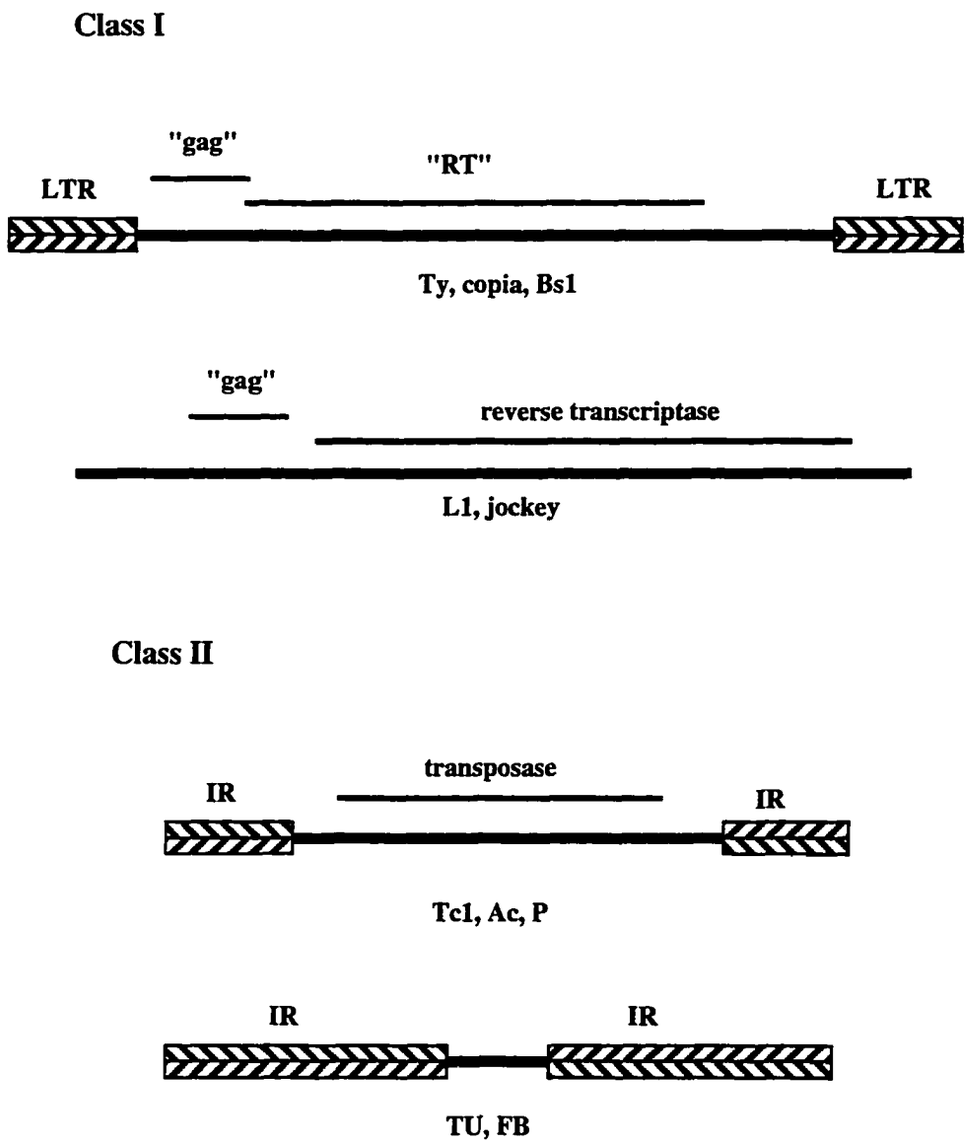


Figure 1.1 Shows the two major classes of transposable elements. The basic structure of the elements is illustrated showing terminal direct (LTRs) and inverted repeats (IRs) as well as coding regions. Several examples of each general type of element are listed below the illustration.

Transposons can generate major chromosomal rearrangements

Evolution of eukaryotic genomes is characterized by numerous changes in chromosome structure. Transposons can increase the frequency at which chromosomal inversions, translocations, deletions and duplications occur. In fact, it was this property of their activity that led to their initial discovery in maize (*Zea mays*) by Barbara McClintock (1948, 1949, 1950). The *Dissociation* element (*Ds*) was initially identified by McClintock as a specific site of chromosome breakage and subsequent rearrangement (reviewed in Fedoroff, 1989). This chromosomal change required the presence of a second element, called *activator* (*Ac*). It was these studies that first lead to the discovery that *Ds* elements could transpose to new chromosomal locations. Since their initial discovery in maize, transposons have been shown to be capable of causing chromosomal rearrangements in many other systems. For example, in *Drosophila melanogaster* P elements are mobilized as the result of a cross between males containing P elements and a female lacking P elements (reviewed in Engels, 1989). The resulting hybrids contain many gross chromosomal rearrangements that are presumably related to the activity of transposons in these individuals (Engels and Preston, 1984). In other situations, rearrangements seem to arise due to recombination between preexisting elements in the genome. In *S. cerevisiae* the breakpoints of numerous deletions, duplications, inversions and translocations contain Ty sequences (Roeder and Fink, 1983), consistent with the idea that they are generated by recombination between elements dispersed throughout the genome. Rearrangements generated by transposons may be the major source of variation in chromosome structure, and an important force in the evolution of the karyotype.

Changes in gene sequences

Many transposons were first identified following their insertion into genes. Insertion of a transposon into a gene can have different consequences depending on where in the gene

the insertion occurs (reviewed in Berg and Howe, 1989). Insertions into coding regions can disrupt gene function. Transposons do not generally contain open reading frames directly at their termini. Therefore, insertions into an exon of a gene can lead to truncation of the gene product due to the introduction of stop codons located near the end of the transposon sequence. For example, insertion of the Tc1 element into the *unc-22* gene of *C elegans* can result in production of an RNA transcript containing element sequence. This leads to translation of a truncated protein product and an *unc-22* mutant phenotype (Moerman et al., 1988). Transposon insertion followed by element excision can lead to more subtle changes in gene sequences. For example, Tc1 elements in *C. elegans* (Eide and Anderson, 1985, 1988; Ruan and Emmons, 1987; Moerman and Waterston, 1991), mariner elements in *Drosophila* (Bryan et al., 1990) and *Mu* elements in maize (Doseff, 1991), are known to leave behind "footprints" after element excision. These footprints vary from single base insertions and deletions to insertion or deletion of many nucleotides (Kiff et al., 1988). Often the footprint contains sequences that were originally part of the transposable element. Footprints generated by Tc1 elements in *C. elegans,* like footprints caused by P elements in *Drosophila* (Gloor et al., 1991), are thought to arise, not from imprecise excision of the element, but from alterations created during the process of DNA repair of the gap left behind when an element excises (Plasterk, 1991). When Tc1 excises it leaves a double stranded break in the DNA. This gap is repaired in a template dependent fashion. Most often, it is the homologous chromosome that is used as a template. In animals homozygous for a Tc1 insertion, the template used to repair the gap will usually be the homologous chromosome, which contains a Tc1 insertion. If repair is precise, no footprint will be observed. If the repair process is interrupted or error prone, it is possible that element or gene sequences near the insertion site will be altered. As described below, transposons also leave "footprints" in RNA sequences when transposon sequences are spliced from RNA transcripts.

Transposons affect splicing of RNA transcripts

Studies in a number of biological systems have demonstrated that transposable element sequences inserted into introns or exons can alter splicing of gene transcripts. This seems to be a feature common to many different elements and is likely to be important in both transposon and genome evolution.

In our lab, Michelle Mills (1993) demonstrated that Tc3 elements can be spliced from *C. elegans unc-22* gene transcripts. Figure 1.2 shows the splicing patterns observed for three different *unc-22* alleles. The *unc-22(r750)* allele contains a 2.3 kb Tc3 insertion in exon 13. Analysis of *unc-22* transcripts from this strain revealed that most of the Tc3 sequence is removed by splicing. A 5' donor site located 40 bases into the end of the Tc3 element is used in conjunction with a 3' acceptor sequence located 11 bp downstream of the Tc3 insertion in exon 13 of *unc-22*. Thus, splicing leaves behind 40 bases of Tc3 sequence and deletes 11 bases of *unc-22* sequence for a net gain of 29 bp. This is a frameshift mutation and results in a *unc-22* loss of function phenotype. The strain harboring the *unc-22 cj213* allele was isolated as a spontaneous wild-type revertant of the strain containing the *r750* Tc3 insertion into *unc-22*. Surprisingly, reversion was found to result from altered splicing of the Tc3 insertion, not element excision. The only difference between the *r750* and *cj213* alleles is the presence of a 4bp insertion at the upstream junction of *unc-22* and Tc3 sequences. Splicing of the *cj213* allele occurs using the same splice sites as the *r750* allele but results in a transcript with a 33bp insertion relative to the wild-type transcript owing to the extra four bases in *cj213*. This alteration leads to the production of an in-frame transcript and a functional gene product. In the case of these *unc-22* alleles, the initial Tc3 insertion is spliced but not in a manner consistent with gene function. A functional gene product is observed only after alteration of the insertion-containing allele.

Many element insertions do not require alteration of sequences to result in wild-type gene function upon splicing of element sequences from gene transcripts. Rushforth and

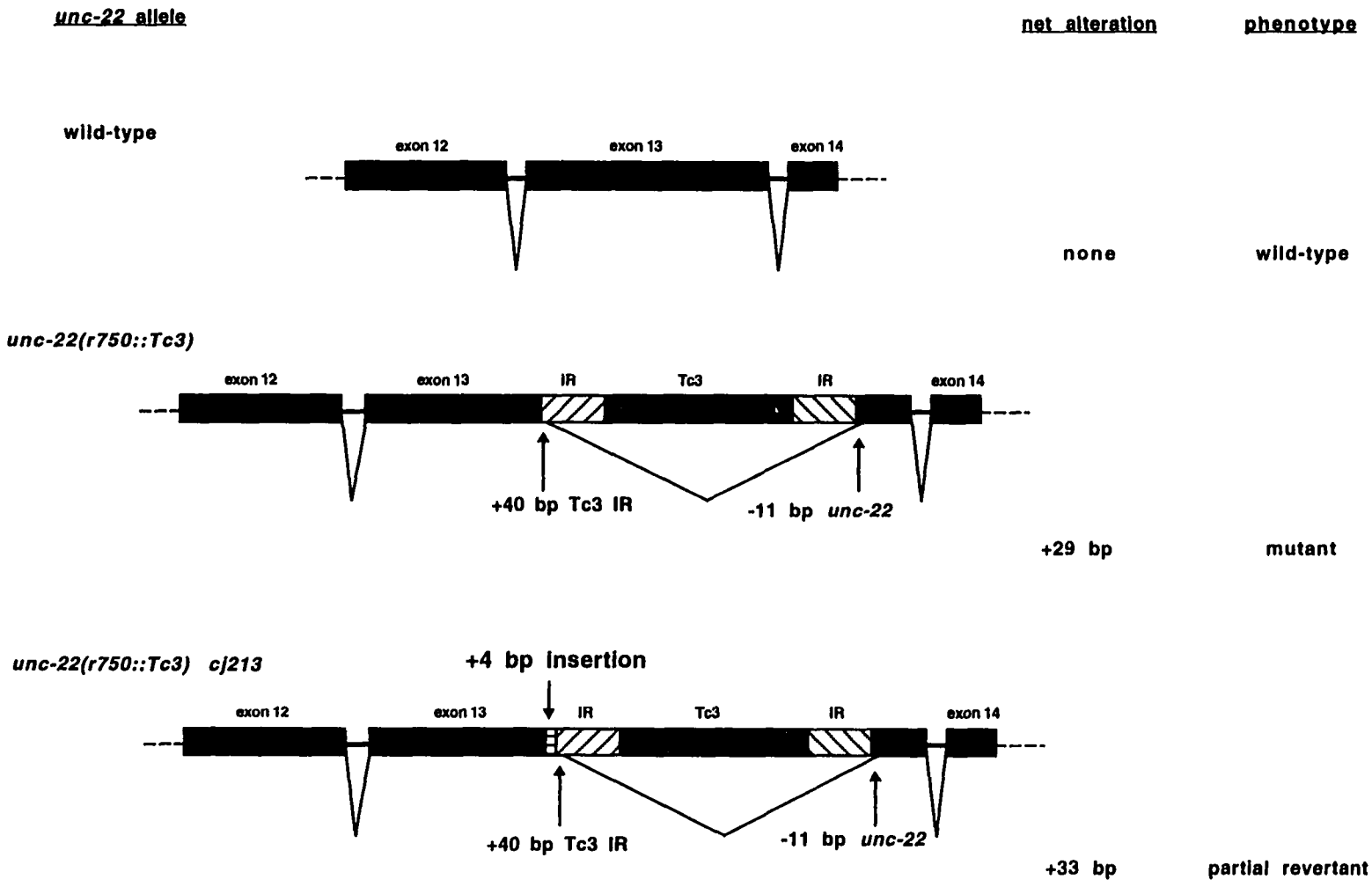


Figure 1.2 Splicing patterns observed for the wild-type unc-22 gene as well as two unc-22::Tc3 alleles.

Anderson (1996) demonstrated that many Tc1 insertions into the *unc-54* and *hlh-1* genes are phenotypically silent due to the splicing of element sequences from gene transcripts. Splicing of transposon sequences has been observed for elements from *Drosophila* (Geyer et al., 1991), maize (Kim et al., 1987; Menssen et al., 1990; reviewed in Wessler, 1989; Purugganan and Wessler, 1992; Purugganan, 1993) and mice (Steinmeyer et al., 1991; Kobayashi et al., 1993). All known transposons lack the sequences required for splicing directly at their termini (although see Menssen, 1990 for a possible exception) . Thus splicing of these elements will invariably lead to alterations in the sequence of RNA transcripts. The number and variety of splice sites used to remove Tc1 sequences from gene transcripts is remarkable (Benian et al., 1993; Rushforth et al., 1993; Rushforth and Anderson, 1996). In some instances splice donor or acceptor sites within the element are used and splicing results in the insertion of portions of Tc1 sequence into transcripts. In other cases cryptic splice sites in the gene are activated upon Tc1 insertion and splicing leads to the deletion of coding sequence. For other Tc1 insertions novel splices sites are used in conjunction with wild-type splice junctions to splice out element sequences. In addition, some insertion-containing alleles can produce several different transcripts when different combinations of alternative splice sites are used to process RNAs encoded by a single gene. Whether splicing results in loss or alteration of gene function will depend on the severity of the change in the RNA transcript and the sensitivity of the product to changes in coding sequence. Altered patterns of splicing induced by element insertion illustrates a potentially significant source of transposon-mediated change in gene sequence and may represent a mechanism for the creation of new introns in genes.

Two transposable element insertions on a chromosome may be capable of mobilizing the sequences contained between them (analogous to composite transposons in bacteria which consist of two insertion sequences flanking a unique region). If these sequences contain open reading frames, and the composite transposable element inserts into the coding region

of another gene, a gene containing a novel exon could be created. Splicing of element sequences from gene transcripts could provide a means of removing element sequences from the gene transcript. This could be a mechanism for exon shuffling, and the creation of genes with new functions (Shapiro, 1992).

Transposon insertions into introns can modify RNA processing patterns by altering host gene splice site choice and creating alternative splicing pathways (Mount et al., 1988; Horowitz and Berg, 1995). Transposons can insert into the sequences required for splicing of an intron leading to inclusion of element and intron sequences in transcripts, and a loss of gene function. Alternatively, activation of cryptic splice sites in the gene or transposon can lead to removal of element and intron sequences and potentially functional transcripts. In at least one instance, a P element insertion in an intron alters transcriptional termination (Horowitz and Berg, 1995). Transposon insertions into introns may alter patterns of splicing, even if they do not disrupt existing splice sites. Even if an intron containing a transposon is spliced using wild-type sites, the presence of element sequences in the unspliced product may alter the efficiency of intron splicing. For some genes the efficiency of intron splicing may affect gene expression. Thus transposon insertions into introns may result in changes in levels of gene expression.

Transposons affect gene regulation

Position effects: Some of the most striking consequences of element activity are their effects on gene regulation. Transposons can increase, decrease, or alter gene expression patterns. One way transposons can alter gene expression is through position effects. I have discussed the role of transposons in generating chromosomal rearrangements, and this is a mechanism that could lead to changes in gene expression. Chromosomal inversion or translocation can result in changes in expression such as when a gene normally found in a euchromatic region of the genome is placed in a heterochromatic region. Alternatively,

chromosomal rearrangement may move a gene into a location where it is placed under the control of regulatory sequences from a different gene can lead to expression under the control of a promoter region from another gene (Schneuwly et al., 1987).

Transposons carry regulatory signals: Transposon insertions occurring in regulatory regions of genes can affect gene expression directly. In the simplest case, insertion disrupts regulatory sequences leading to a decrease in gene expression. In other cases transposable element insertions bring a gene under the control of a different set of regulatory signals. Changes in gene expression induced by insertion of Ty elements in *S. cerevisiae* provide an interesting example of the phenomenon (Errede et al., 1987). ROAM (Regulated Overproducing Alleles responding to Mating type) mutations result from insertion of Ty elements into the 5' flanking region of genes. Expression of these alleles is increased relative to wild type and surprisingly, transcription of ROAM alleles is regulated by mating type (prior to insertion of Ty, these genes do not respond to mating type) . It is known that levels of expression of Ty element encoded products is significantly lower in diploids than in haploids. The ROAM alleles acquire their novel response to mating type because of cis-acting elements present in Ty sequences inserted into gene regulatory regions.

In plants, the two-element systems in maize are the best characterized transposons. These elements can lead to different changes in gene expression of an affected locus depending on what other elements are present in the genome. For example, insertion of the nonautonomous receptor element (*Rs*) into the *Bz* locus conditions normal pigmentation of mature kernels due to splicing of element sequences from gene transcripts (Kim et al. 1987). In the presence of an autonomous *Spm* element elsewhere in the genome, however, gene expression from the *Bz* locus is suppressed leading to a loss of pigmentation (Klein and Nelson, 1983).

Many transposable elements contain open reading frames and promoters and enhancers

that regulate their expression. In fact, almost every sequence motif known to play a role in controlling gene expression can be found within transposable elements (McDonald, 1990). Thus, insertion of transposons into gene regulatory regions and subsequent alteration of gene expression patterns may be a common feature of transposons. As I discuss in the next section, the elements themselves often respond to particular regulatory pathways and may confer novel tissue specific or developmental patterns of gene expression on genes near their site of insertion. Transposition is the only mutational mechanism known to generate such specific changes in gene regulation. Because of this unique attribute, it is likely that transposons are a major source of regulatory variation in genetic systems. It is precisely this sort of variation that is thought to play a critical role in macroevolutionary change (Britten and Davidson, 1969; Wilson et al. 1974).

Regulation of transposable element activity:

I have described some of the mutagenic properties of transposable elements. One obvious feature of this activity is the diversity of mutations caused by transposons. An additional feature, which I have not discussed, is the abundance of mutations that are transposon-induced. In *Drosophila*, where frequencies of spontaneous mutation have been estimated for many element families, it is thought that at least half of all spontaneous mutations are due to the activity of transposons (Green, 1988). Many spontaneous mutations are deleterious and it is expected that unchecked transposition would be extremely detrimental to an individual. Therefore it is not surprising that mechanisms exist to regulate when, where and how transposons move.

In multicellular organisms transposable elements can be regulated in a tissue specific manner. In *C. elegans*, transposon activity is regulated by tissue specific factors. Collins et al. (1987) screened 1,500 EMS-mutagenized animals for elevated reversion frequencies of an *unc-54*::Tc1 mutant. They isolated several strains with reversion frequencies that

were as much as 100-fold higher than the parental strain. Reversion occurs from element excision from the *unc-54* locus in the germline. Somatic excision of Tc1 from the *unc-54* locus was also examined and found to occur at levels comparable to the parental strain. This indicates that excision of Tc1 responds to different regulatory signals in the germline and the soma. Transposons in maize and *Drosophila* also show evidence for tissue specific regulation (reviewed in Berg and Howe, 1989).

Some transposable element activities correlate with developmental stage. Maize *Ac/Ds* elements are developmentally regulated. *Ac/Ds* element excision events occurring early in the development of a tissue give rise to large sectors of revertant tissue. However, increasing the number of *Ac* elements in the genome leads to excision later during tissue development and smaller patches of revertant tissue (McClintock, 1948; Schwartz, 1984). Interestingly , this effect (a copy number-dependent delay in timing of excision) occurs in different tissues regardless of the number of cell divisions that have elapsed. Increasing *Ac* copy number results in smaller patches of revertant cells (i.e. a delay in timing of excision) among different tissues within the same plant. Hence, excision seems to relate to the physiological state of the cell and is somehow related to the number of remaining divisions in a cell lineage.

Tissue-specific and developmental regulation of transposon activity suggest that element activity can respond to host encoded factors. As described above, Ty elements in yeast can lead to changes in gene regulation, such as ROAM mutants. These mutations occur because Ty elements respond to host encoded factors. Subsequent studies demonstrate that many yeast genes are required for proper transcription of Ty elements (Boeke et al., 1989). The relationship between levels of Ty mRNA and levels of transposition are still largely unknown.

Transposons also respond to environmental conditions, probably mediated through host factors. For example, mutator elements in maize are responsive to ultra-violet light

(Walbot, 1992). Ty elements also respond to UV light, and in addition have been shown to be activated by DNA damaging chemicals and gamma irradiation (Morawetz, 1987; McEntee and Bradshaw, 1989). Thermal stress in *Drosophila* leads to transcription of *Drosophila* heat shock genes as well as copia retrotransposons (Junakovic et al., 1986).

Element encoded factors may also play a role in keeping transposon activity in check. P elements in *Drosophila* encode a transposase protein as well as a repressor of P activity (Engels, 1989). In, fact the repressor is likely encoded in the same gene as the transposase (Handler et al., 1993). It is possible that the repressor functions by out-competing transposase for binding sites in element sequences. Alternatively, since transposases often function as multimers, truncated or altered products of the transposase gene might disrupt the transposase complex. It appears that regulation of transposon activity is complex and controlled by a combination of host and element encoded factors.

The selection of sites for transposable element insertion is another case where element and host-encoded factors interact to regulate element copy number and distribution. Variable levels and degrees of insertion site specificity are observed for all transposons where preference for insertion site have been examined. In the most extreme cases element insertion is restricted to a single target sequence. The R2Bm element in insects (a non-LTR retrotransposon) always inserts into the same site within one of the many copies of a the rRNA genes (Luan et al., 1993). Other transposons insert preferentially into different regions of the genome. For example, *Mu1*-related elements in maize show little preference for specific target sequences, but preferentially insert into sequences that are present in low copy number in the genome (Cresse, et al., 1995). Maize *Ac* elements show a preference for insertion sites linked to the donor site (Dooner and Belachew, 1989; Schwartz, 1989) as do P elements in *Drosophila* (Tower et al., 1993). Ty elements in yeast show a preference for insertion into regions containing tRNA genes, LTRs , or previously inserted transposable elements (Ji, et al., 1993). In *C. elegans*, target site preference within a single

gene (*gpa-2*) was examined for the related transposons Tc1 and Tc3 (van Luenen and Plasterk, 1993). Both elements insert exclusively into a TA dinucleotide. Some sites in the gene are hotspots for insertion whereas other potential insertion sites, often located only a few nucleotides away from a hotspot, are not used at all. Other than the absolutely conserved TA at the site of insertion, no other significant consensus for insertion was observed for either Tc1 or Tc3. Surprisingly, the distribution of insertion sites was very different for the two related elements suggesting that Tc1 and Tc3 recognize different features of the target DNA when inserting. The evolution of insertion site preferences is complicated because it involves coevolution of element and host sequences. Additionally, the distribution of element insertions is filtered by natural selection, so it may be difficult to determine if an observed distribution arises as the result of insertion site preference or natural selection.

Transposons and evolution:

I have discussed in some detail the mechanisms by which transposons introduce genetic variation and how element activity is regulated. I would now like to discuss the evolutionary implications of these issues. Two basic sets of questions are of interest to those of us studying transposons. How do the elements themselves evolve? and How does element activity affect the evolution of the genomes that contain them? These questions are difficult to address because we can only observe transposons present in the genomes of extant organisms and must make inferences about how they evolved. It is possible to study the distribution of element sequences to gain an understanding of the forces acting on element sequences and the potential role these elements have played in gene and genome evolution. Studying the biochemical nature of transposition and its regulation may also provide important insights. However, since we can never know exactly how evolution occurred, we are often forced to explain the current state of transposons in the genome in

terms of the selective forces that have determined their structure, distribution, and effect on the genome. In discussing these ideas I think it is important to make a distinction between "levels of selection" to explain the perspective from which we should view element evolution.

Levels of Selection

Most of us are familiar with the concept of phenotypic selection. This implies that the unit of selection is an individual. This is the ordinary means by which natural selection is thought to occur. Alleles increase in frequency if they enhance the fitness of the their bearers relative to that of genetically different individuals in the same population. This concept often leads to the conclusion that perhaps the only way a particular DNA sequence can ensure its survival in the genome is by ensuring the survival of the individual it inhabits (Doolittle and Sapienza, 1980). With respect to transposons we can see how this type of selection may be important. If a particular transposon insertion leads to a deleterious phenotype, the individual harboring such a mutation may be eliminated by natural selection. Selection acting at the level of individual organisms may explain the evolution of mechanisms that repress transposition since individuals which keep transposons in check may be more fit than individuals with higher levels of activity. However, in other cases we must consider selection acting at other levels.

Group selection is a concept often evoked to explain the evolution of traits, such as altruism, that appear to provide no particular advantage to an individual, but may be advantageous to a group of organisms (such as the species as a whole). Group selection arguments are sometimes cited in discussions of transposable element evolution. Transposon activity can be a significant source of genetic variation. Under some conditions this variation may be beneficial. This has lead some to argue that transposons persist because they are advantageous to their host species (Campbell, 1983). One

advantage that they could provide is a source of potentially adaptive genetic variation that could be useful to an organism during periods of environmental stress (Arnault and Dufournel, 1994). This is analogous to saying that transposons serve a function, as a storehouse of potentially advantageous genetic variation. Proponents of such a view cite examples of increases in the rates of transposition and excision during periods of genomic shock or under harsh environmental conditions. Although compelling, group selection arguments are often criticized because of the difficulty in changing allele frequencies among groups (Williams, 1966). Since there are by definition many more individuals than there are groups containing these individuals, natural selection can act much more quickly to alter allele frequencies among individuals than among groups. Since most transposition events within an individual are expected to be neutral or deleterious, selection against transposons is likely to occur. It seems unlikely that selection favoring transposons because of their advantage under rare periods of environmental stress is strong enough to overcome the persistent selection against transposons within individuals. However, just because group selection arguments should be viewed cautiously, does not mean that they do not describe some attributes of transposon evolution.

Both phenotypic selection and group selection rely on the concept that if a sequence is present in the genome, it must have a function, even if it is not obviously apparent. This often leads to the creation of elaborate adaptive stories to explain element and genome evolution. As Gould and Lewontin (1979) write, "the rejection of one adaptive story often leads to its replacement by another, rather than to a suspicion that a different kind of explanation might be required. Since the range of adaptive stories is as wide as our minds are fertile, new stories can always be postulated". The concept of genic selection may provide such a 'different kind of explanation'.

Genic selection is perhaps the most useful concept in understanding transposon evolution. Genic selection involves selection at the level of the genes. Within an

organism, DNA sequences that multiply in the germline will be overrepresented in the next generation. Provided that multiplication is not too deleterious for the host, elements that replicate more efficiently will have a selective advantage over elements that replicate slowly. This concept of genic selection lead to the term selfish DNA to describe transposable elements (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). Selfish DNA is DNA that replicates along with the host DNA, but has no function. As Futumya (1986) writes "transposable elements persist in spite of their effects on organisms, not because of them". That is to say, their only function is their own self-preservation. Under the selfish DNA hypothesis, transposons are expected to evolve more efficient means of replication until they reach a point where they are detrimental to their host, and natural selection (phenotypic selection) favors loss of the offending elements. Thus, there is no reason to expect transposons to reach a point of equilibrium where copy number is stable over time. There may be a constant battle for element survival within a cell. Genic selection may initially favor elements that replicate efficiently. If these elements reach a point were their activity is harmful to their hosts, selection will favor regulation of element activity or loss of the elements.

The evolution of transposon sequences

The ultimate origin of transposable elements is unknown and is likely to remain a mystery. Some would argue the evolution of selfish DNA sequences is inevitable (Doolittle and Sapienza, 1980). If a mutation arises that increases the probability of survival of a particular DNA sequence, and that mutation has no effect on the phenotype of the organism, it will persist by genic selection. So, transposons are likely to be ancient cellular inhabitants. Given the selfish nature of their replication, the only precondition for the evolution of transposons is the existence of machinery capable of replicating them. It is very unlikely that there was a single origin for transposons. In particular, the DNA

transposons (class I) and the retrotransposons (class II) almost surely have independent origins.

A hypothesis for the relationships among class I retrotransposons has been developed based on comparisons of reverse transcriptase genes contained within the elements (Xiong and Eickbush, 1990; reviewed in McDonald, 1993). It is thought that bacterial retrons are the most ancient type of retroelement. Rooting the reverse transcriptase tree along this branch reveals that the non-LTR elements are the progenitors of the LTR elements. This is also consistent with the proviral hypothesis (Temin, 1980) that states that retroviruses evolved from cellular retrotransposons. Another group of retrotransposons are called SINEs. They rely on reverse transcription to transpose, but do not encode the enzyme themselves. SINEs, such as Alu elements in humans, are thought to be derived from reverse transcription of cellular RNA polymerase III transcripts (Okada, 1991).

Class II, or DNA transposons, share some structural features. They all contain inverted repeat sequences at their termini, and many contain one or more open reading frames, at least one of which encodes a transposase. Transposase genes, unlike reverse transcriptase genes of retrotransposons, are often very different from each other. For this reason the relationships among different families of class II elements is ambiguous. Attempts to align the amino acid sequences of different transposases have revealed several motifs that suggest a relationship between elements found in species ranging from bacteria, to *C. elegans*, to *Drosophila*, and fish (Doak et al., 1994; Henikoff, 1992). It is also interesting to note that most of these elements insert into TA dinucleotides. Others have noted similarities in the transposase genes between elements in plants and *Drosophila* and suggest a common origin for these transposons (Calui, 1991). Reconstructing the relationships among transposable elements is complicated by departures from strict vertical transmission of transposon sequences.

Almost every discussion of the evolution of transposons concludes that some of the

relationships among elements found in different organisms arise from horizontal transmission of elements. These arguments suggest that element sequences may be capable of crossing species boundaries. Evidence for these claims comes primarily from comparisons of element sequences from different taxa. In some cases almost identical elements are found in distantly related species, whereas more closely related species do not share elements with a similar sequence (Robertson, 1993). In these circumstances, transfer of element sequences between species is invoked as an explanation for their taxonomic distribution. The vectors that mediate horizontal transfer of transposons are unknown. Viruses with the ability to infect a broad range of hosts (such as insect baculoviruses) have been implicated as possible vectors (Miller and Miller, 1982). Parasitic mites that infect diverse taxa are also potential vectors (Houck et al., 1991). Some have noted the similarity in the structure of a mites mouth parts to laboratory microinjection needles that are used to introduce foreign DNA into laboratory organisms (McDonald, 1993).

Many other factors influence the fate of transposon sequences in a genome. In addition to transposition and excision, transposons may be targets for recombination and gene conversion. These processes can lead to the phenomenon of concerted evolution which has been used to explain the greater than expected similarity in sequence between members of a multigene family (Hartl and Clark, 1989). Gene conversion is one mechanism that can lead to homogenization of sequences within a multigene family across the genome (Walsh, 1987). Unequal crossing over among tandemly repeated sequences can have a similar effect. Transposons, like multigene families, may be subject to concerted evolution.

The evolution of genomes containing transposons

I have described the multitude of mutational effects generated by transposons, their ubiquitous phylogenetic distribution, and the significant contribution of transposons to spontaneous mutation. One of the major questions remaining is: What role have

transposons played in genome evolution? Most of the time we are forced to speculate on this role, since the remnants of transposon activity in the genome are likely to decay quickly. Occasionally, researchers identify cases where transposon sequences seem to unambiguously accompany the evolution of a new function. Perhaps the best known case is that of the mouse *Slp* gene which encodes the sex-limited protein (Stavenhagen and Robins, 1988). The *Slp* gene is part of the murine histocompatibility complex and is believed to have arisen from a tandem duplication of another gene called *C4*. The genes share significant sequence similarity, however they show very different patterns of tissue-specific expression. Characterization of the cis-regulatory sequences responsible for the different patterns of gene expression revealed that the enhancer sequence that confers androgen responsiveness on the *Slp* gene, but not the *C4* gene, is contained within the LTR of a cryptic retroviral like element (Stavenhagen and Robins, 1988). The insertion appears to be ancient since the element contains numerous substitutions within the LTRs as well as a number of frameshifts and nonsense mutations within the coding region of the reverse transcriptase gene of the retroviral element. As genome sequencing projects progress we are sure to find many more "smoking guns", where the remnants of transposon mediated alterations in the genome are plain to see.

Methods for understanding the madness:

Transposons have been described in a large number of species. However, studies of their biological activity and evolution have been pursued in a handful of model systems Several general approaches for investigating their behavior and evolution are described below.

A. Biochemical analysis has focused primarily on understanding the molecular basis for transposition. Determination of the factors required for transposon activity and their

interactions with regulatory molecules and transposon sequences are of particular interest. At his point the characterization of transposase function at the biochemical level exists for a few systems. *In vitro* transposition systems have been developed for a few systems (Mizuuchi, 1983; Morisato and Kleckner, 1987; Kaufman and Rio, 1992). The first and best characterized is phage Mu, a bacteriophage with transposon like activity. The active form of the Mu transposase is a tetramer that is formed only in the presence of element sequences (Baker and Mizuuchi, 1992). For other systems, *in vitro* transposition systems have not been developed and biochemical dissection of the components necessary for transposition are more difficult. In *C. elegans* the polypeptide encoded by the transposon Tc1 has been investigated *in vivo* and *in vitro* (Schukkink and Plasterk, 1990; Vos et al., 1993). The Tc1 transposase (known as Tc1A) is a DNA binding protein that binds specifically to sequences within the Tc1 element. Nuclear extracts from strains overexpressing Tc1A were used in gel retardation assays with labeled portions of the Tc1 inverted repeat sequence. These studies suggest that Tc1A and probably other factors form a complex that mediates Tc1 transposition. Similar studies of the polypeptide encoded by the transposon Tc3 indicate that it binds specifically to the sequences within the Tc3 inverted repeats (van Luenen et al., 1993).

B. Geneticists have approached the same questions as biochemists using different methodology. Most genetic approaches begin by identifying new insertions as spontaneous mutations that alter the expression of a gene leading to a visible phenotype. Subsequently, mutations are isolated that enhance, suppress or alter the phenotype of the original insertion. In some cases these mutations can be used to identify genes that are involved in the regulation of transposon activity. As described above, these techniques have lead to the identification of numerous genes in *S. cerevisiae* that control transcription of Ty elements. Genetic methods have also proven useful for estimating the rate of

transposon insertion and excision. Insertion into a gene can be monitored by screening for transposon induced mutants, and excision events can be examined by screening for revertant animals (e.g. Eide and Anderson, 1988). These topics will be addressed in greater detail in Chapter III. Genetic methods in *C. elegans* have allowed the identification of several loci which increase the frequency of transposition and excision (Collins et al., 1987). To date, none of these genes has been cloned and the basis for their control of transposon activity remains a mystery.

C. Experimental evolution can be used to simulate transposon evolution in the laboratory. These types of experiment are used to address questions such as: Are transposons a burden to their hosts? Do they ever provide a selective advantage? and What are the fates of element sequences upon introduction to a naive genome? These experiments have been limited to a few organisms that can be cultured under controlled conditions. P element transposition in Drosophila has been shown to contribute substantial new variation for the quantitative trait abdominal bristle number (Torkamanzehi, et al., 1992). Experiments with yeast (Wilke, et al., 1993) and bacteria (Hartl and Dykhuizen, 1984; Chao and McBrown, 1985; Hall, 1988; Modi et al., 1992) have demonstrated that transposons can provide a selective advantage to their hosts . However, these effects may in part be due to the culture conditions (often organisms grown in chemostats) and may not reflect the action of selection in natural populations.

D. Genome level analysis of transposon sequence and distribution have been used to understand the evolutionary dynamics of transposons in natural populations. Most studies of this type have been carried out using *D. melanogaster*. Charlesworth and Langley (1989) studied the population frequencies of transposons at chromosomal sites by means of *in situ* hybridization of transposon probes to polytene chromosomes. In general, they

found that transposon insertions were present at very low frequencies at individual nucleotide sites from *Drosophila* population samples. The exact nature of the forces responsible for these distributions is still unclear, but the theoretical predictions suggest that selection may act to reduce the likelihood of recombination between elements located in different regions of the genome (Charlesworth et al., 1992). Researchers have used genetic and molecular biological techniques to examine the phylogenetic distribution of elements and the distribution of sites within a genome and between individuals in natural populations of *Drosophila.*

Transposons in *C. elegans*

C. elegans as a model

The nematode *C. elegans* has emerged as one of the premier model organisms used to understand the process of development and elucidate the molecular basis of animal behavior. Several features of *C. elegans* makes it an ideal system for molecular genetic analysis. *C. elegans* is a small (<1mm long, 959 cells), transparent, free-living nematode. It reproduces as a self-fertile hermaphrodite and produces brood sizes of approximately 300 animals. Males arise spontaneously at low frequency in natural populations, and can be maintained as stocks in the lab for performing genetic crosses. Thousands of animals can be cultured on a single petri dish containing an agar media, and *E. coli* as food. Molecular and genetic methods are routine in this organism and have provided much insight into the molecular mechanisms controlling development.

Two additional resources available to the *C. elegans* research community distinguish this nematode from other model systems. The first is the fate map constructed for the *C. elegans* cell lineage. *C. elegans* is the only metazoan where the entire series of cell divisions, from the fertilized egg to the mature adult, have been determined. The fate of every cell in the organism is known and the process of development is essentially invariant

between individuals. This information has proved to be invaluable in studies of the mechanisms controlling development. Thousands of mutants have been identified with altered patterns of development. This has lead to the characterization of many genes controlling cell differentiation, determination, and even cell death. In addition, laser microsurgical techniques are available that allow the perturbation of individual cells. This has allowed scientists to investigate the interaction of cells during development. In addition to a cell fate map, *C .elegans* will be the first multicellular organism to have the complete nucleotide sequence of its genome determined. As of June 1996, almost 70% of the 100Mb genome has been sequenced (Bob Waterston, personal communication) although only about 30Mb of the sequence is available in Genbank. The complete sequence may be available by 1998.

Many of the genes controlling development and behavior of *C. elegans* have already been identified. The challenge is to understand how these genes interact with each other to give rise to a mature functioning animal. Once the complete nucleotide sequence of the genome has been determined, attention will focus on assigning a role to the thousands of genes whose function is unknown. This avenue of investigation requires the use of reverse genetic approaches. The term reverse genetics is applied to techniques used to target mutations to loci whose function is in question. In the yeast *S. cerevisiae*, mutations are conveniently targeted to specific loci by homologous recombination. In *C. elegans*, homologous recombination has not been developed as a tool to introduce mutations. Instead, reverse genetic methods in *C. elegans* have relied on transposable elements to generate specific mutations. As a consequence of their ability to move and generate mutations, transposons are used extensively as tools for introducing specific genetic alterations. As we come to understand the molecular basis for transposition and the regulation of element activity, we will be able to improve and simplify the use of transposons as tools for reverse genetic approaches.

Thesis organization

The thesis that follows is divided into three sections that represent related phases of my investigations of transposable elements in *C. elegans*.

There are many interesting questions regarding transposon sequence evolution within a genome. Chapter II contains a description of transposable elements identified in *C. elegans* followed by my investigation of variation in DNA sequences, transposase sequences and genomic location among transposable elements in the *C. elegans* genome. In addition to understanding the evolution of transposon sequences in a genome it is important to understand the phenotypic consequences of transposon activity. Chapter III describes my attempts to use molecular techniques to investigate the phenotypic consequences of Tc1 insertion. The consequences of insertion are dependent on when and where transposition occurs. Chapter IV describes the characterization of tissue-specific and developmentally regulated patterns of Tc1 activity.

CHAPTER II

TRANSPOSONS IN THE *C. ELEGANS* GENOME: VARIATION WITHIN AND BETWEEN ELEMENT FAMILIES

Introduction:

Discovery of transposons in *C. elegans*

Tc1 was the first transposable element described in *C. elegans*. It was identified as the source of multiple restriction length polymorphisms between two common laboratory strains of *C. elegans*, Bristol and Bergerac. Several restriction fragments, 1.6 kb larger in Bergerac than in Bristol, were identified by Southern hybridization with unique sequence probes (Emmons et al., 1979). Comparison of these restriction fragments demonstrated that the 1.6 kb size difference was due to the presence of a repeated sequence element that was dispersed throughout the genome, and present at about 30 copies in Bristol and 300 copies in Bergerac (Emmons et al., 1983; Liao et al., 1983). The first Tc1 element sequenced was from the Bergerac strain (Rosenzweig et al., 1983). It is 1610 bp long with 54 bp perfect terminal inverted repeats (IRs) and contains two ORFs, the larger of which could encode a 273 amino acid polypeptide. All known Tc1 insertions occur into the dinucleotide TA and duplicate the target site upon insertion. Tc1 has traditionally been described as showing remarkable sequence conservation between different copies of the element in the genome. However, some heterogeneity between elements has been described (Rose et al., 1985; Harris and Rose, 1989). Most of the Tc1 elements in Bristol and Bergerac appear to be the same length although some restriction sites differ between elements. At least 4 different enzymes reveal differences among Tc1 elements. Restriction analysis of 17 cloned Tc1 elements from the Bristol genome shows that 1 has a 55bp

insert, 2 have 700bp deletions and at least two have single-base polymorphisms (Moerman and Waterston, 1989).

After the discovery of Tc1, several other families of transposable elements were identified in the *C. elegans* genome. The Tc2 element was serendipidously discovered as an IR containing sequence located within a clone containing a Tc1 element (Levitt and Emmons, 1989). The first Tc2 element sequenced (Ruvolo et al., 1992) was 2074 bp long with 24 bp perfect terminal IRs. In addition to IRs, Tc2 contains degenerate subterminal direct repeats that are arranged in a complex overlapping pattern. Tc2 elements contain 3 ORFs capable of encoding a polypeptide. The number of copies of Tc2 varies from approximately 4-25 between strains. Individual copies of Tc2 were cloned from genomic libraries of Bristol and Bergerac. In contrast to Tc1 which showed little variation between elements, restriction mapping of the Tc2 elements contained in these clones revealed significant restriction site variation between elements (Levitt and Emmons, 1989).

Tc3, Tc4, and Tc5 elements were all identified as new insertions into genes isolated in the *mut-2* strain TR679. As described in chapter I, *mut-2* mutants have a greater level of Tc1 activity than wild-type strains. In addition to Tc1, *mut-2* mobilizes Tc3, Tc4 and Tc5 element families. Tc1 elements move in several genetic backgrounds that lack the *mut-2* mutation but germ-line activity of Tc3, Tc4, and Tc5 has not been detected in any genetic background lacking the *mut-2* mutator. So it is possible that these three elements are not active at all in wild-type genetic backgrounds. However, it is known that Tc3 elements are capable of movement in a Bristol background when a Tc3 transposase gene driven by an inducible promoter is overexpressed in transgenic animals (van Luenen et al., 1993; Vos et al., 1993).

Tc3 was isolated as a new insertion into the *unc-22* gene (Collins et al. 1989). Tc3 is 2335 bp long with 471bp terminal IRs. It contains 2 ORFs capable of encoding a 329 amino acid polypeptide. Tc3 always inserts into the dinucleotide TA and duplicates this

target sequence upon insertion. There are 12-18 copies of Tc3 among various strains and restriction digestion reveals little size heterogeneity among Tc3 elements (Collins et al., 1989). The Tc3 transposase shows some similarity to the polypeptide encoded by Tc1.

Tc4 was identified as the cause of a mutation in the *ced-4* gene (Yuan et al., 1991). Tc4 is 1605 bp long with 774 bp terminal IRs. This structure has been referred to as a fold-back element since the sequence consists of almost entirely IR. Tc4 does not contain any significant ORFs, although a "variant" Tc4 element called Tc4v does contain an ORF (Li and Shaw, 1993). All Tc4 insertion sites examined occur into a pentanucleotide sequence CTNAG. The central trinucleotide TNA is duplicated upon insertion. Copy number seems to be about 20 among several strains. Like Tc2 elements, restriction analysis revealed significant heterogeneity among different copies of Tc4.

Tc5 was discovered as a new insertion in the *unc-22* gene (Collins and Anderson, 1994). Tc5 is 3171bp long with 491bp terminal IRs. It contains several ORFs capable of encoding a 532 amino acid polypeptide. Tc5 also inserts into the pentanucleotide CTNAG and duplicates the central trinucleotide TNA upon insertion. The number of copies of Tc5 varies from 4 to 7 between different wild-type strains.

Tc6, like Tc1, was identified as the cause of a restriction length polymorphism between Bristol and Bergerac strains (Dreyfus and Emmons, 1991). One Tc6 element is 1603 bp, contains 765 bp IRs and does not have a large ORF. Like Tc4, Tc6 has the structure of a foldback element. To date there is no direct evidence for Tc6 transposition. Only the polymorphisms between strains due to the presence of Tc6-like sequences argues for the ability of these elements to transpose. Sequencing of two additional Tc6 elements or partial elements revealed the presence of at least one deleted copy and one copy with complicated rearrangements in the Bristol genome.

Distribution of element insertion sites across the genome

Theoretically, we might expect a transposable element to increase in copy number until all available sites are occupied (Ajioka and Hartl, 1989). In reality, the pattern that has emerged from studies of several *Drosophila* elements is that most target sites are occupied at low frequency in a population (Montgomery and Langley, 1983; Ronsseray and Anxolabehere, 1987). These observations have led some to conclude that elements are maintained by a transpositional increase in copy number but are kept in check by one or more opposing forces (Charlesworth et al., 1992). The frequency of sites occupied on the X chromosome has lead Charlesworth et al. (1992) to suggest that element frequencies are higher for sites that experience lower rates of recombination. This may be due to selection acting against elements that could participate in ectopic exchange (homologous recombination between elements at nonhomologous locations in the genome).

Analysis of transposable elements in the *C. elegans* genome will provide a different perspective on transposable element and genome evolution. The complete nucleotide sequence of the Bristol genome will allow characterization of all transposable elements in a single genome. The Bristol strain is distinguished by having an extremely low level of transposable element activity. Germ-line insertion and excision of Tc1, Tc2, Tc3, Tc4 and Tc5 elements is undetectable in this strain (with the exception of one Bristol subline which acquired Tc1 mutator activity; Babity et al., 1990). Reproduction in *C. elegans* occurs mainly by self-fertilization, so individuals within a population of Bristol animals can be considered essentially isogenic with respect to their transposable element copy number and distribution. Therefore, the sites containing transposons in the Bristol genome can be described as "resident sites", that is, sites that are stably inherited in the strain. These resident sites arise as the product of the transposition process that distributed the elements across the genome, and selection or genetic drift which lead to their current distribution.

None of the *C. elegans* transposons characterized to date have long consensus

sequences for insertion. Insertion site preferences are best studied for Tc1 and Tc3 elements which both insert into the dinucleotide TA. Mori et al. (1988) and Eide and Anderson (1988) proposed similar consenesus sequences for Tc1 insertion based on 16 independent Tc1 insertions. However a larger dataset of 204 independent Tc1 insertions and 166 independent Tc3 insertions (van Luenen and Plasterk, 1994) reveals that other than the absolute requirement for TA, there is no other strong consensus for insertion site for either element. Considering that the *C. elegans* genome is AT-rich, there are likely to be an extremely large number of sites with a primary sequence suitable for insertion. Tc1 and Tc3 elements are each represented by fewer than 30 copies in the Bristol genome and therefore represent a very small fraction of the potential insertion sites.

Transposon-like sequences in the *C. elegans* genome sequence

Analysis of the first 2.2 Mb of contiguous sequence from the *C. elegans* genome revealed some interesting inverted repeat containing sequences (Wilson et al., 1994). Only sequences with inverted repeats less than 1kb apart and at least 70% identical between IRs were considered. Most are small with, on average, IRs of 70bp with 164bp of internal unique sequence. These sequences are found approximately once every 5.5kb in the contig, but their distribution in the genome is nonrandom. 43% of the repeats occur in introns which account for only about 20% of the total sequence. It has been suggested that these small IR containing sequences can be clustered into families, but the similarity among different elements in a family has not been described.

Oosumi, Garlick and Belknap (1995) describe methods of computational analysis to identify inverted repeat domains in DNA sequences. They have applied their methods to identify other element sequences in the *C. elegans* genome including sequences with similarity to Tc1, Tc2, Tc5 and mariner transposons (W.R. Belknap, personal communication). Initial results came from analysis of 2.2 Mb of genome sequence

(Oosumi et al., 1995) in which they describe many elements that share similarity to the ends of Tc2 elements. One sequence in particular, a ~345bp element called Cele2 , was repeated 36 times within the 2.2 Mb contig. This one element alone accounts for almost 1% of the total 2.2Mb of sequence. Oosumi et al. (1995), suggest that these elements that are similar to known transposons at their termini, but are generally shorter, and are nonautonomous elements analogous to those described in maize.

McClintock (1950) distinguished autonomous copies of a transposon, which were able to move on their own, from nonautonomous elements that can move only in the presence of an autonomous element. At the sequence level, the difference between autonomous elements and nonautonomous elements can often be traced to differences in one or more of the coding regions of the element. Nonautonomous elements frequently contain multiple substitutions, deletions, or insertions that disrupt the coding region. Apparently, many of these modified elements are still recognized by the transposase and can be mobilized in trans by other elements in the genome. For example, autonomous P elements in Drosophila are 2907 bp long, but shorter nonautonomous elements are also found in the Drosophila genome (Spradling and Rubin, 1982). One nonautonomous P element accounts for over half of the copies of P in some natural populations (Black et al., 1987). This variant P element contains a large 1753 bp internal deletion, but still contains 31bp IRs and encodes a truncated protein product that could act as a repressor of P element transposition. This illustrates an important point regarding nonautonomous elements. Even though they do not encode the factors necessary for their own transposition, they may play important roles in regulating the activity of both autonomous and nonautonomous elements in the genome.

The relationships between different families of transposable elements in the *C. elegans* genome as well as the relationships within some element families are still unresolved. With the large amount of data available from the sequencing project (about 30 Mb thus far) it is

difficult to identify all of the transposon-like sequences, let alone characterize the relationships among different families. I contribute to the description of *C. elegans* transposons in the genome by comparing sequences which resemble known transposons in *C. elegans*. I began by using BLAST (Altschul et al., 1990) to identify cosmids containing transposon-like sequences. The genomic location of each cosmid was determined and compared to the position of other elements. Each element sequence was examined for IRs, and when identified, the IRs were compared to determine the degree of similarity between the ends of the element. The transposon-like sequences were aligned to the previously described transposons, and to each other. Where possible, the relationship between element sequences was determined. These analyses reveal both remarkable similarities as well as differences between these transposon sequences and contributes to our understanding of transposable element sequence evolution within a genome.

Methods:

As of June, 1996, the Genbank database contains approximately 30 Mb of *C. elegans* sequence from several linkage groups. This represents close to one-third of the total *C. elegans* genome (100Mb).

The Genbank database was searched using the National Center for Biotechnology Information (NCBI) BLAST (Altschul et al., 1990) server. The entire transposable element sequences for Tc1, Tc2, Tc3, Tc4, Tc5, and Tc6 were used as search queries. For each element, I chose to examine the top ten (or so) sequences which showed greatest similarity to the known element. Ten sequences were generally enough to identify several copies of the known transposon as well as copies of additional sequences from related element families.

The sequences from Genbank that I chose to examine came entirely from the cosmids

that were sequenced as part of the *C. elegans* genome project (i.e. they are all from the Bristol strain). Cosmid clones are given a unique identifier by the sequencing consortium. This consists of a letter followed by an additional 3-6 letters or numbers (e.g. c28f5). In rare cases this convention is not used (e.g. cosmid Ac3). Throughout this discussion I use the cosmid name to refer to the transposon-like sequence found within a particular cosmid.

Each of these cosmids has been ordered into large contigs. The genomic location of each cosmid was determined using ACeDB (A *C. elegans* Data Base; Thierry-Meig and Durbin, 1992). Each cosmid has been fingerprinted in order to determine overlaps between cosmids and generate the large contigs (Coulson et al. 1986). Within a contig, each clone is given a position in terms of a range of pMap (physical map) values. The length unit of the *C. elegans* physical map is the fingerprint band. Although fingerprint bands are not strictly physical measures, on average a band is about 1.83 kb (Barnes et al., 1995). I determined the pMap positions for all of the cosmids which contained transposon-like sequences and used them to generate a map of transposon sequences in the Bristol genome.

Cosmids contain inserts of approximately 30 kb. The transposon-like sequences were extracted from the larger cosmid sequences (using the EDITSEQ module of DNA*, copywrite DNASTAR, Inc.) for further analysis. Initially, the putative ends of an element contained on a cosmid were identified by determining where cosmid sequences matched the ends of the known transposon in the alignment generated during the BLAST search. Since the ends of a "new" element could be longer than predicted by these criteria, approximately 10 additional nucleotides were retained at both ends of every element sequence extracted from a cosmid sequence. These extra bases were also retained to examine the sequences flanking the element insertions.

All of the known *C. elegans* transposons contain terminal IRs. To examine the length and structure of IRs within the transposon-like sequences, each element was reverse complemented and aligned to itself. These pairwise alignments were performed using the

GAP program with the GCG package of programs (Devereux et al., 1984). The number of nucleotide changes between the inverted repeats of a single element as well as the number of gaps were determined for all sequences. In a few cases elements were analyzed using dotplots generated within the ALIGN module of DNA*.

Multiple sequence alignments were performed on subsets of the sequences which shared identity based on the BLAST results and preliminary alignments. These alignments were generated using the PILEUP program within GCG (Devereux et al., 1984). Gap and gap length penalties were adjusted to maximize the number of paired bases in the alignment. In the cases of elements with similar terminal IRs, but great differences in element length, gap penalties had to be significantly reduced. Whenever possible the full length element sequences were aligned. In a few cases internal regions of a long transposable element had to be deleted to accomplish proper alignment of element termini. These deleted elements contain the cosmid name followed by the three letters "del". In addition, the elements are not all located on the same strand of the cosmids. Therefore, care had to be taken to align elements in the correct orientation relative to other sequences (not always a simple task with sequences that contain long IRs). Sequences which were reverse complemented relative to the strand submitted to Genbank as the "+" strand, contain the cosmid identifier followed by the two letters "rc".

To further examine the relationship between sequences, I used PAUP version 3.1 (Phylogenetic Analysis Using Parsimony, Swofford, 1993). PAUP allows convenient inclusion and exclusion of characters and taxa from an alignment. I used PAUP to build trees from entire elements as well as conserved portions of elements to aid in the description of relationships among different copies of transposon-like sequences. Trees were bootstrapped to determine the statistical significance of groupings. In no case is there an element sequence that represents an obvious outgroup for the tree. Ideally, the choice of an appropriate outgroup would depend on knowledge of the time of divergence among

elements and the relationship of *C. elegans* transposons to elements in closely related nematodes. Since this kind of information is currently unavailable for most of the elements, all trees were midpoint rooted. Midpoint rooting places the root at the center of the longest branch in the tree. Assuming that substitutions between elements accumulate in a clock-like manner, this method should separate the most divergent sequences and provide at least a first estimate of the historical relationships among sequences.

Results and Discussion:

Genomic Distribution:

Table 2.1 contains the names of all the cosmid sequences analyzed in this study followed by their genomic location in terms of contig (ctg) and pMap value (see methods). The pMap values in Table 2.1 were used to generate Figure 2.1 which shows the distribution of transposon and putative transposon sequences in the *C. elegans* genome.

At this time only portions of some chromosomes have been sequenced. The breakdown of sequence by linkage group (LG) is approximately (as stated in a progress report from the *C. elegans* genome Consortium):

LGI <1 Mb, LGII 7.2 Mb, LGIII 7.2 Mb, LGIV 3.9 Mb, LGV <1 Mb, LGX 11.8 Mb.

Note that sequences of LGIII and LGX are nearly complete, whereas sequencing of LGI and LGV has just begun.

The genomic locations of all of the transposon and transposon-like sequences considered in this study are shown in Figure 2.1. It is important to consider that only portions of the genome have been sequenced and even the best covered regions contain gaps. In addition, initial sequencing efforts have focused on the gene dense regions of chromosomes (the central regions; Barnes, 1995). Therefore, at this time, a detailed statistical analysis of element distribution is premature. There do appear to be several clusters of elements in Figure 2.1. For example, there are several elements in a fairly small

Table 2.1: Genomic location of transposon-like sequences in the *C. elegans* genome. The location of cosmid clones on the *C. elegans* physical map is given in terms of pMap values (Coulson, 1986).

element type	cosmid	pMap(lower)	pMap(upper)	contig	chromosome
5	c01b7	625	647	313	V
1	zk856	1382	1404	313	V
6	ac3	1507	1534	313	V
6	f53b7	1816	1833	313	V
1	r03h10	-3507	-3477	369	II
2	k03h9	-2162	-2148	369	II
1	f18c5	-2109	-2092	369	II
5	t13c2	-1988	-1972	369	II
5	f31e8	-1978	-1959	369	II
1	c07d10	-1693	-1676	369	II
3	r10h1	-1630	-1605	369	II
1	c28f5	-1610	-1581	369	II
6	zk669	-1370	-1346	369	II
3	f27e5	-126	-108	369	II
5	zk930	1008	1039	369	II
2	t10f2	-2083	-2068	377	III
2	k10d2	-2075	-2064	377	III
6	zc395	-2020	-1998	377	III
6	f48e8	-1912	-1885	377	III
6	w03a3	-1711	-1696	377	III
2	f01f1	-1681	-1644	377	III
4	zk686	-666	-644	377	III
4	c27d11	-643	-623	377	III
5	f44b9	-549	-519	377	III
3	b0303	-161	-134	377	III

Table 2.1 continued: Genomic location of transposon-like sequences in the *C. elegans* genome.

element type	cosmid	pMap(lower)	pMap(upper)	contig	chromosome
5	C48b4	318	346	377	III
6	zk180	-2530	-2511	423	IV
3	t13a10	-1603	-1593	423	IV
6	c33h5	-967	-933	423	IV
2	t26a8	-603	-577	423	IV
6	t26a8	-603	-577	423	IV
2	f56d5	-87	-58	423	IV
1	zk1251	111	130	423	IV
2	zk792	1224	1244	423	IV
3	c25g4	1679	1697	423	IV
4	f49e11	2057	2072	423	IV
5	c04e7	-2501	-2486	674	X
5	t19d7	-2357	-2338	674	X
2	f53h8	-2161	-2141	674	X
4	r04b3	-1190	-1172	674	X
2	f52b10	-946	-925	674	X
3	k10b3	-816	-782	674	X
4	f32a6	289	313	674	X
2	c15b12	937	957	674	X
4	f23g4	995	1018	674	X
5	c39d10	1637	1655	674	X
3	zc64	2088	2103	674	X
1	m02d8	2101	2121	674	X
1	d1009	2188	2205	674	X
5	c24a3	2219	2245	674	X
1	zk899	2607	2620	674	X
1	f08g12	3636	3660	674	X
2	zk455	3654	3680	674	X

Table 2.1 continued: Genomic location of transposon-like sequences in the *C. elegans* genome.

element type	cosmid	pMap(lower)	pMap(upper)	contig	chromosome
1	r173	3742	3760	674	X
4	f57g12	4076	4092	674	X
1	f19h6	4183	4205	674	X
5	t14g8	4456	4479	674	X
1	f02d10	4781	4806	674	X
3	zk1086	4983	5007	674	X
4	f23c11	6409	6422	674	X
1	f23a7	6493	6507	674	X
1	c30g4	6916	6943	674	X
4	t08g2	7013	7040	674	X
3	t25g12	7030	7048	674	X
1	f10d7	7064	7087	674	X

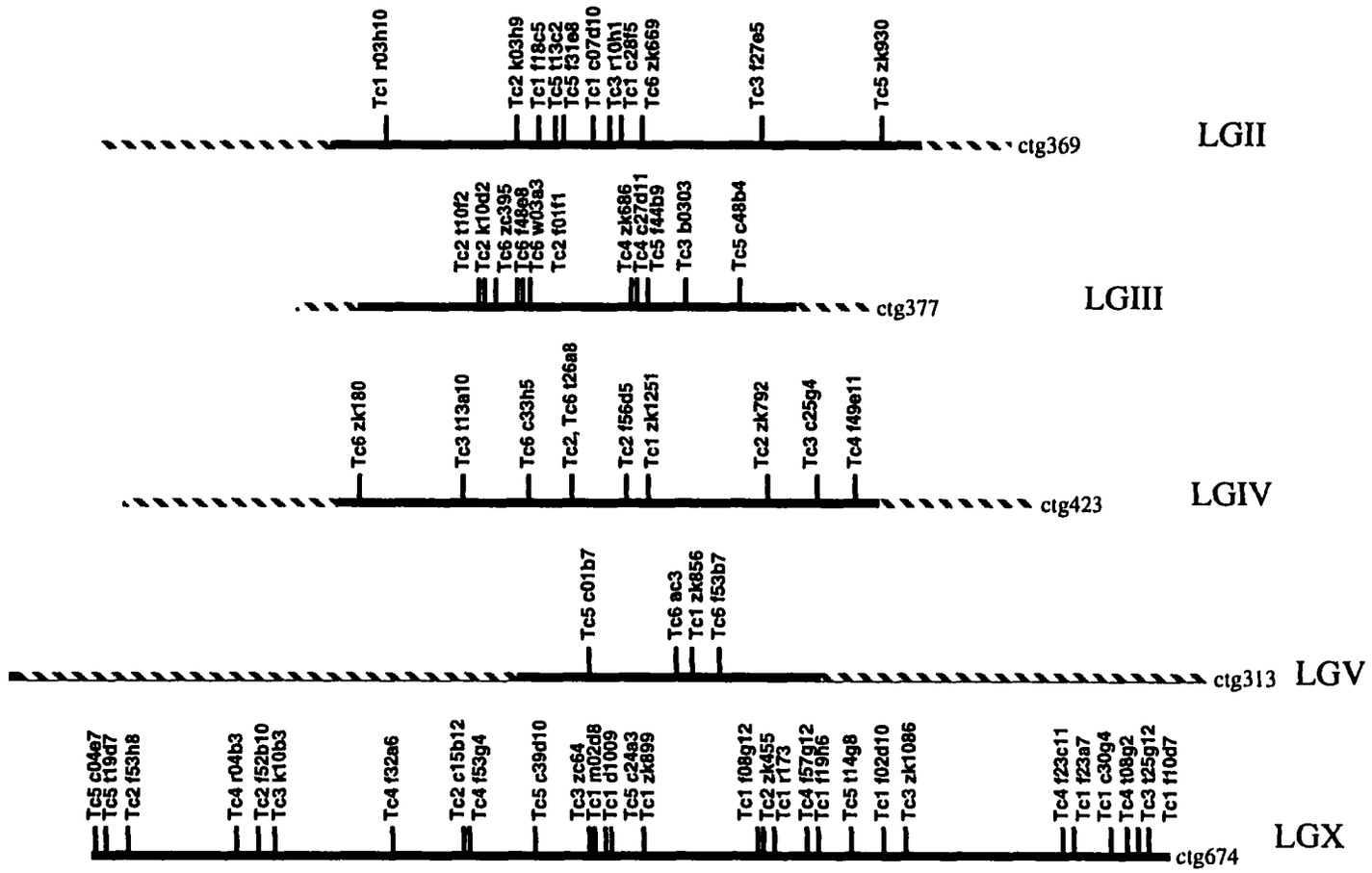


Figure 2.1 Diagram showing the position of transposon-like sequences on the major contig for each chromosome. BLAST hits are indicated above the contigs. The label includes the name of the element used as a BLAST query followed by the name of the cosmid containing the sequence. Stipled lines within the contigs indicate regions of the genome that have not been sequenced. Note that not all contigs or chromosomes are shown in this diagram since some have not been sequenced and hence no transposon-like sequences were identified.

region of LGIII. In addition there are eight cases where two elements of the same group (e.g. two Tc1-related elements) occur in close proximity to each other. Although there are not a large number of copies of any element represented, in many cases similar elements are found on different LGs. This suggests at least one interesting feature of all of the sequences used in this study: they are all found at dispersed locations in the genome. This is one of the hallmarks of a transposable element. So, based on the similarity to known transposons identified in the BLAST analysis and the genomic location of these sequences, I predict that these sequences are transposons or transposon-derived sequences.

Analysis of Tc1 and Tc1-like sequences in the *C. elegans* genome:

The 14 cosmid sequences with the highest scores after a BLAST search with the entire Tc1 sequence were compared to each other and to the canonical Tc1 element sequenced from the Bergerac strain. Significant features of their structure are summarized in Table 2.2. Of the 14 sequences there are:

Seven elements with perfect 54 bp IRs like those found in Tc1

Six of which are approximately the same length (1610-1611 bp) as Tc1.

One is considerably shorter (929 bp).

Six with 348-349 bp IR and total lengths ranging from 878-923 bp.

These IR are not perfect (26-34 sites vary within the IRs of each element).

One sequence with 276 bp IR (19 sites vary within the IRs) that is 804 bp long

I compared Tc1 to one of the 6 elements with 348 bp IRs in a dotplot analysis and observed no long segments of identity between the two elements. The only region where they are obviously similar is over the 38 bp at each end of the element. In fact , the elements are identical at 36 out of 38 nucleotides at each end. It was this region of identity which was identified by the BLAST search with Tc1. The one element with 276 bp IR appears to match the first 276 bp of these six elements with long IRs. The dataset was

Table 2.2: Comparison of the length of IR, variation among the two IR of each element, and total length for Tc1 and cosmid sequences of high-scoring BLAST hits.

cosmid	IR (bp)	variable sites between IR	indels (bp)	length (bp)
Tc1	54	0	0	1610
ZK1251	54	0	0	1611
R03H10	54	0	0	929
ZK856	54	0	0	1610
C28F5	54	0	0	1611
F18C5	54	0	0	1611
R173	54	0	0	1611
F08G12	54	0	0	1611
C30G4	276	19	1,1	804
F02D10	348	34	1	922
C07D10	348	32	1	921
F19H6	348	26	43,1	878
ZK899	348	32	1	921
D1009	349	32	1,1	923
M02D8	349	32	1,1	923

divided in half at this point and the elements were studied in two groups, one consisting of Tc1 elements, the other containing the elements with long IRs.

Appendix A contains an alignment of of the seven Tc1 elements to the canonical Tc1 element isolated as a RFLP between Bristol and Begerac strains (Rosenzweig et al, 1983). All seven Tc1 elements contain identical, perfect 54 bp terminal IRs. One element contained in cosmid r03h10, contains a 682 bp internal deletion relative to the other elements. None of the sequences are identical.

All elements differ from the published sequence in one respect, they contain an extra T at position 361 relative to Tc1 (this was noted by others examining Tc1 elements in Bristol). This base may be important since it brings a potential ATG start codon in frame with a putative upstream ORF that allows Tc1 to encode a 343 amino acid transposase (without the extra base, only 273 amino acids are predicted). The only additional size variation between these Tc1 elements is a single base deletion in zk856 at position 198 in the alignment, in a 4 bp polyT run (upstream of the coding region).

Table 2.3 contains a distance matrix showing the number of pairwise differences between Tc1 sequences in the alignment shown in Appendix A. None of the sequences

Table 2.3: Pairwise distances between Tc1 and cosmid sequences for positions 11-1621 of the APPENDIX A alignment. Absolute distance are shown in the lower diagonal. Mean distances (adjusted for missing data) are shown in the upper diagonal.

	1	2	3	4	5	6	7	8
1 C28f5rc	-	0.004	0.001	0.004	0.004	0.003	0.006	0.002
2 Tc1	6	-	0.004	0.004	0.004	0.003	0.006	0.001
3 F18c5rc	2	6	-	0.004	0.004	0.003	0.006	0.002
4 R173rc	6	6	6	-	0.004	0.003	0.006	0.003
5 Zk1251	6	6	6	6	-	0.003	0.006	0.001
6 Zk856rc	5	5	5	5	5	-	0.005	0.002
7 F08g12rc	9	9	9	9	9	8	-	0.004
8 R03h10	2	1	2	3	1	2	4	-

differ by more than 9 out of the 1611 bases in the alignment. Surprisingly, of the few changes observed, many occur within the open reading frames. Table 2.4 shows the

Table 2.4: Variable sites from an alignment of predicted transposases for Tc1 and seven Tc1-like elements. Each column heading indicates the ORF (1 or 2) followed by the position of the amino acid residue in that ORF. Sequences that match C28f5rc are indicated by a quotation mark. Gaps are shown as "."

	1.26	1.30	2.40	2.114	2.120	2.174	2.211	2.212	2.213	2.215	2.243	2.279	2.281
C28f5rc	I	M	M	S2	V	L	R	R	R	H	I	Q	V
Tc1	.	.	"	"	"	F	"	"	H	"	"	"	"
F18c5rc	"	"	"	G	"	"	"	"	"	"	"	"	"
R173rc	"	"	"	"	L	"	"	"	"	"	"	"	F
Zk1251	"	"	"	"	"	"	"	"	"	R	V	"	"
Zk856rc	"	"	"	"	"	"	"	"	"	"	"	"	"
F08g12rc	T	"	T	"	"	"	P	C	"	"	"	L	"
R03h10	"	"	"	"	"	.	.	.	.	.	.	.	.

associated amino acid replacements. There are 13 variable sites. Four changes are observed within a 4 aa stretch of the protein. There are no shared amino acid polymorphisms between these sequences.

r03h10, the Tc1 element containing a 682 bp deletion, could encode a 184 aa polypeptide that is identical to the full length Tc1 protein over the first 178 amino acids and contains six additional aa's that are encoded from a region of Tc1 that does not usually contain an ORF.

Figure 2.2 shows a Tc1 tree inferred by parsimony derived from the complete element sequences from the alignment in Appendix A. It is bootstrap consensus tree and is midpoint rooted. c28f5rc and f18c5rc are more similar to each other than to any other sequence and f08g12rc is the most divergent. Gaps were not informative in this analysis since no gaps were shared between sequences.

Appendix B contains an alignment of the seven remaining BLAST hits containing IRs with similarity to Tc1. The IRs of these elements are much larger and more variable than those found in Tc1 (see Table 2.2). Several insertions and deletions (indels) were observed between elements. f19h6 has a 44 bp deletion relative to the other sequences. c30g4 is the most divergent sequence. c30g4 is smaller than the rest but has 276 bp IRs like the terminal 276 bp of the other elements. On one side c30g4 contains sequences that are similar to the rest of the 348 bp of the IRs in the larger elements. The central region of the c30g4 element aligns poorly with these other elements. c07d10 and zk899 share a 1 bp deletion at position 521and d1009 and m02d8 share a 1 bp insertion relative to the other sequences at position 839.

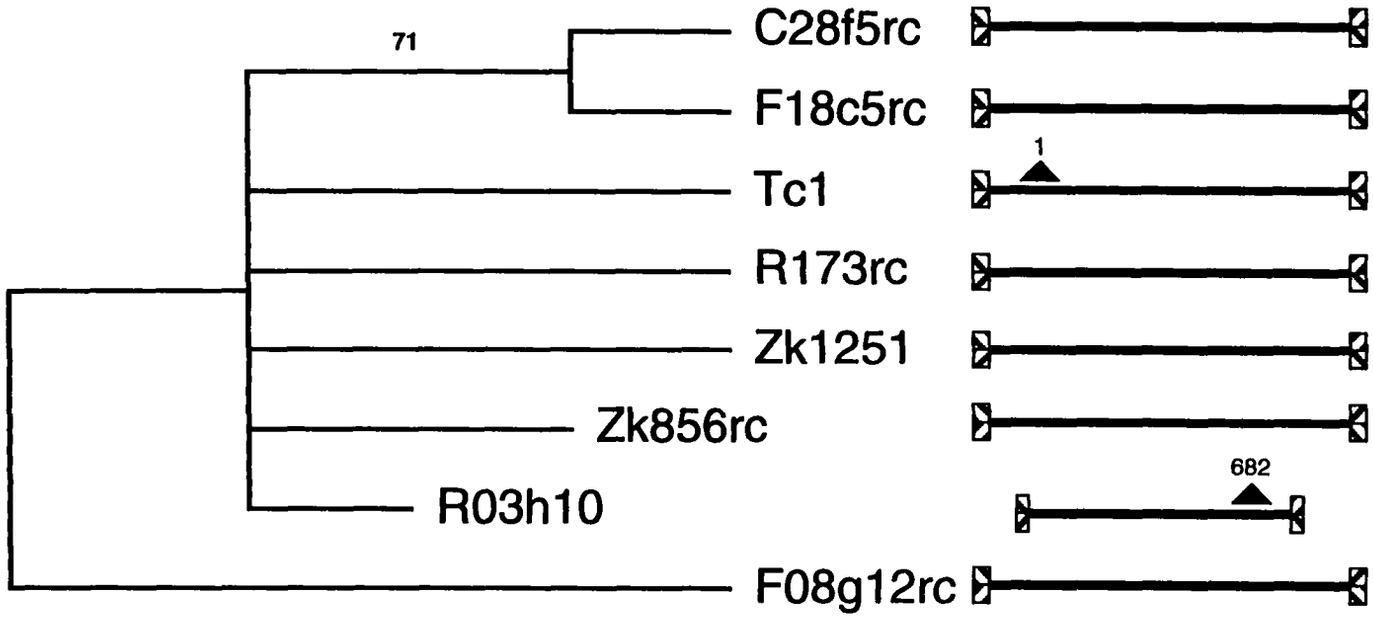


Figure 2.2: Parsimony bootstrap consensus tree (100 replicates) of 8 Tc1 elements. The tree was constructed using the entire element sequences, positions 11-1621 in the alignment in Appendix A. The diagrams to the right of the tree show some of the major structural features of each element. Striped regions indicate IRs and insertions and deletions are indicated by an arrowhead with a number above it indicating the length of the indel (arrows pointing up are deletions and arrows pointing down are insertions).

Table 2.5 contains a distance matrix for sequences aligned in Appendix B. Excluding c30g4, no 2 sequences differ at more than 4 sites. This implies that there are far more differences between the IRs of a single element than there are differences over the whole length of separate copies of the element. Many of the differences between the IRs are conserved between elements suggesting that the changes occurred prior to transposition of these elements.

Table 2.5: Pairwise distances between Tc1-like cosmid sequences for positions 11-936 of the APPENDIX B alignment. Absolute distance are shown in the lower diagonal. Mean distances (adjusted for missing data) are shown in the upper diagonal.

	1	2	3	4	5	6	7
1 C07d10	-	0.000	0.002	0.002	0.004	0.005	0.234
2 Zk899	0	-	0.002	0.002	0.004	0.005	0.234
3 D1009rc	2	2	-	0.000	0.002	0.002	0.238
4 M02d8	2	2	0	-	0.002	0.002	0.238
5 F02d10	4	4	2	2	-	0.005	0.240
6 F19h6	4	4	2	2	4	-	0.241
7 C30g4	187	187	191	191	192	182	-

The differences in the IR are scattered, the first change occurs within the first 30 bases of the element. There are some single base indels between IRs and one large deletion in f19h6. All of these elements have the structure of a foldback element with 348 bp IRs and 226 bp in the middle.

Figure 2.3 shows a tree illustrating the relationships among the Tc-1 like elements with a fold-back structure. They cluster into two well supported groups, separating c07d10 and zk899 from the rest.

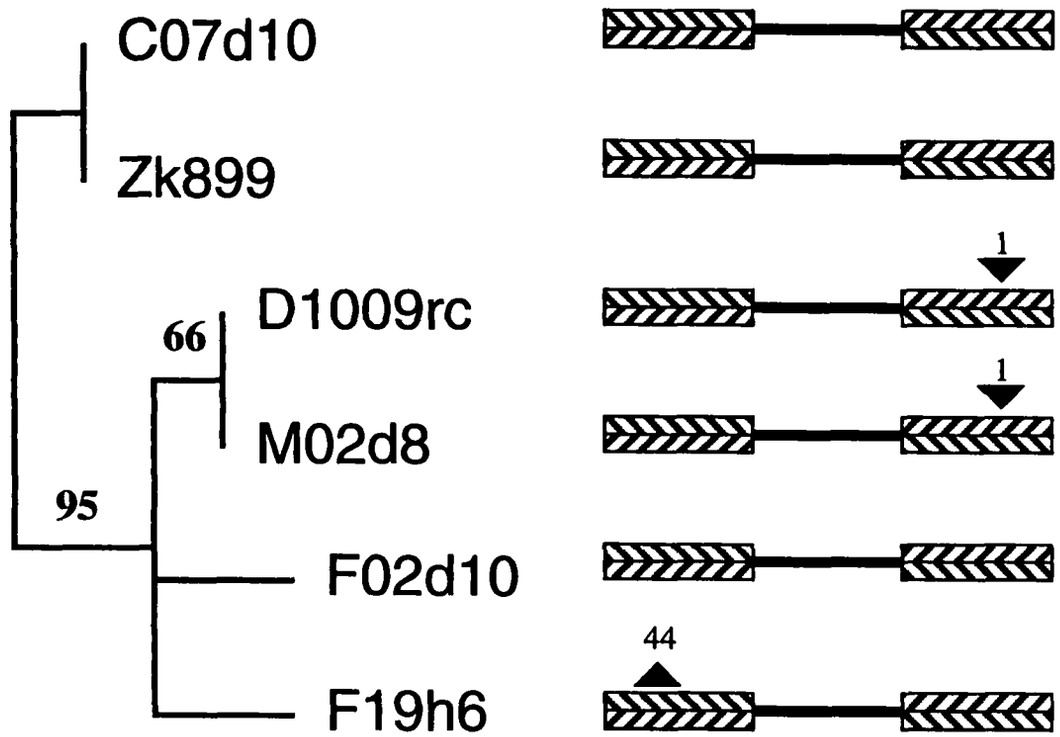


Figure 2.3 Parsimony bootstrap consensus tree (100 replicates) of foldback Tc1-like elements based on positions 11-936 of the alignment in Appendix B. The diagrams to the right of the tree show some of the major structural features of each element. Striped regions indicate IRs and insertions and deletions are indicated by an arrowhead with a number above it indicating the length of the indel (arrows pointing up are deletions and arrows pointing down are insertions).

Analysis of Tc2 and Tc2-like sequences in the *C. elegans* genome:

The 10 cosmid sequences with the highest scores after a BLAST search with the entire Tc2 sequence were compared to each other and to the canonical Tc2 element sequenced from the Bergerac strain. Significant features of their structure are summarized in table 2.6. Of the 10 sequences, clearly none of the elements is much like the canonical Tc2

Table 2.6: Comparison of the length of IR, variation among the two IR of each element, and total length for Tc2 and Tc2-like cosmid sequences.

cosmid	IR (bp)	variable sites between IR	indels (bp)	length (bp)
Tc2	24	0	0	2074
zk455	26	0	0	466
f01f1	25	0	0	446
f52b10	26	5	0	446
f53h8	26	0	0	431
t26a8	26	0	0	425
zk792	26	1	0	413
t10f2	26	0	0	421
k03h9	-	-	-	427
f56d5	26	0	0	424
c15b12	26	10	0	445

element. They are much shorter than Tc2 ranging from 413 to 466 bp in size (compared to the 2074 bp Tc2 element). All have IRs of approximately 26 nt, the same size as Tc2 IRs. Most of the IRs are perfect. k03h9 is similar to the other elements at one end, but has a 3' terminal deletion relative to the other elements and therefore lacks IRs altogether. c15b12

has several substitutions in its left IR.

Appendix C contains an alignment of Tc2-like elements with a modified Tc2 sequence (Tc2del). The Tc2 sequence was modified to improve the alignment of the ends of the elements and contains a large internal deletion in the middle of the element. Dotplots clearly indicate that there is no significant similarity between the central ~1800 nts from Tc2 and the Tc2-like elements. The Tc2-like elements and Tc2 have 26 bp IRs and share similarity over approximately the first 130 bp and last 110 bp. There is clearly a repetitive structure within this region. Short (~18 nt) sequences are repeated approximately 4 or 5 times in this short region. The repeat begins within the IR. The number of copies of repeat differs between elements. Copies of the repeat within an element are interupted by other sequences, mostly polynucleotide runs.

There is lots of variation between these elements, including size variation. Lots of small indels are found, some of which are shared between elements. Table 2.7 contains a list of sites which show length variation between sequences. Note that in the alignment in Appendix C at position 121 all sequences have a 35bp deletion relative to Tc2 and all similarity to Tc2 breaks down at this point in the alignment

Table 2.7: Lists the position of gaps found among Tc2 related sequences in the alignment shown in Appendix C. Note that only gaps found in more than one sequence are included in the table. No one sequence served as a reference for determining the presence of insertions and deletions (indels).

position in alignment	indel	contained in elements
52	+1	f56d5, t26a8, t10f2, zk792
81	-2	f56d5, t26a8, t10f2, zk792
121	+19	f53h8, zk455
234	-2	f01f1, f52b10, k03h9
284	-1	f53h8, zk455
330	-21	f56d5, t26a8, t10f2, zk792

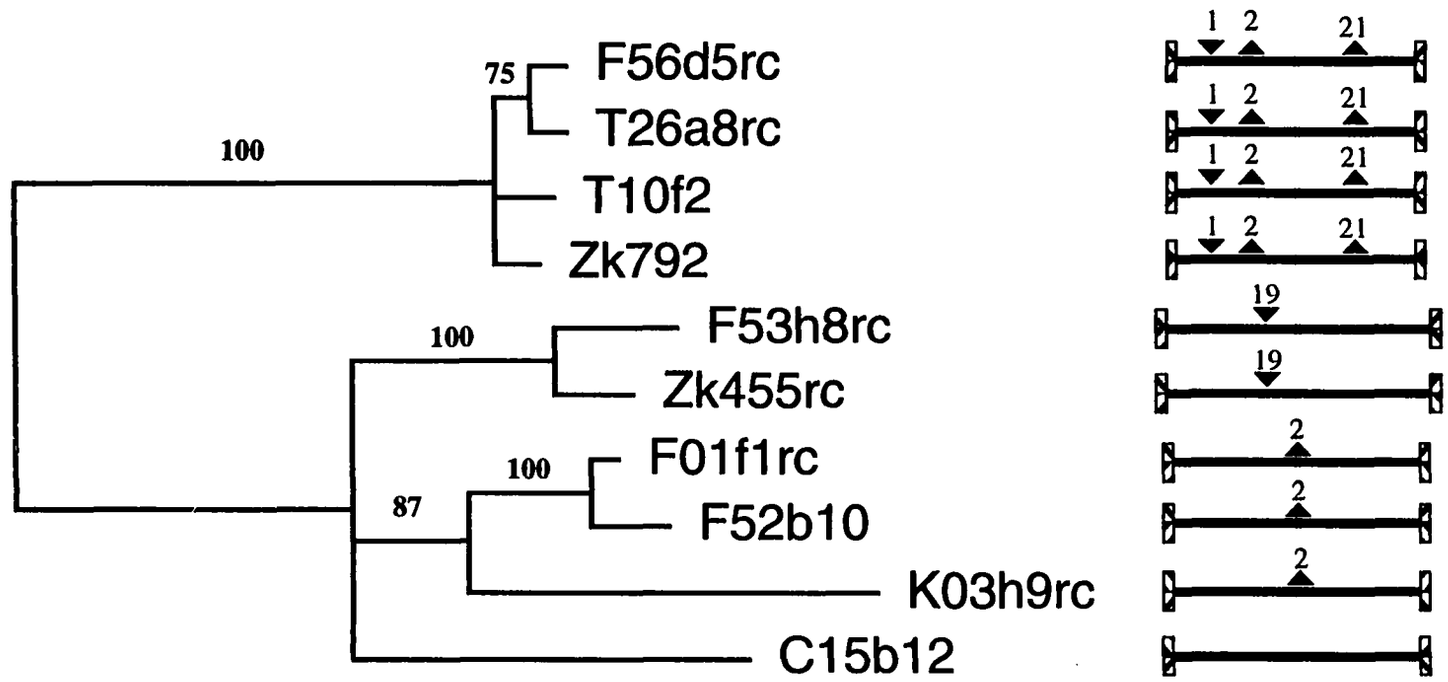


Figure 2.4 Parsimony bootstrap consensus tree (100 replicates) of Tc2-like elements based on positions 11-499 in the alignment shown in Appendix C. The diagrams to the right of the tree show some of the major structural features of each element. Striped regions indicate IRs and insertions and deletions are indicated by an arrowhead with a number above it indicating the length of the indel (arrows pointing up are deletions and arrows pointing down are insertions).

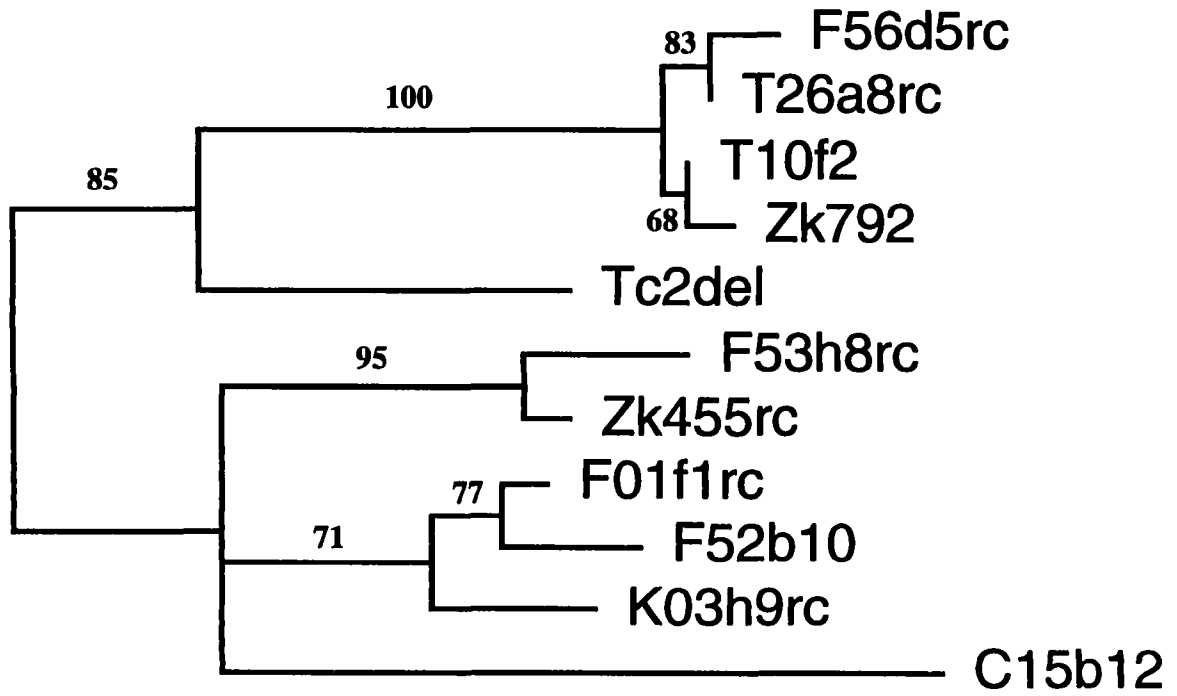


Figure 2.5 Parsimony bootstrap consensus tree (100 replicates) of Tc2del and related elements based on positions 11-120 and 391-499 from the alignment in Appendix C

Table 2.8: Pairwise distances between Tc2-like cosmid sequences for positions 11-499 of the APPENDIX C alignment. Absolute distance are shown in the lower diagonal. Mean distances (adjusted for missing data) are shown in the upper diagonal.

		1	2	3	4	5	6	7	8
1	F56d5rc	-	0.024	0.043	0.031	0.318	0.325	0.306	0.299
2	T26a8rc	10	-	0.040	0.031	0.312	0.314	0.296	0.289
3	T10f2	18	17	-	0.032	0.312	0.313	0.297	0.289
4	Zk792	13	13	13	-	0.314	0.314	0.298	0.290
5	F53h8rc	123	121	120	118	-	0.065	0.146	0.166
6	Zk455rc	137	133	131	129	28	-	0.146	0.157
7	F01f1rc	129	125	124	122	60	65	-	0.034
8	F52b10	126	122	121	119	68	70	15	-
9	K03h9rc	118	116	117	113	50	58	36	39
10	C15b12	149	147	147	144	83	88	87	88

		9	10
1	F56d5rc	0.330	0.353
2	T26a8rc	0.323	0.348
3	T10f2	0.330	0.351
4	Zk792	0.325	0.350
5	F53h8rc	0.144	0.203
6	Zk455rc	0.152	0.198
7	F01f1rc	0.094	0.196
8	F52b10	0.102	0.199
9	K03h9rc	-	0.189
10	C15b12	72	-

Figure 2.4 shows a midpoint rooted bootstrap tree illustrating the relationships among "full length" Tc2-like elements without Tc2. Two distinct clusters of sequences are well supported. The clustering of sequences in the tree based on nucleotide sequence differences across the whole elements is consistent with the distribution of shared gaps among sequences. There is a lot of variation among these Tc2-like sequences, which allows for good resolution of the relationships among these sequences.

Figure 2.5 shows a tree built using the sequences at the end of all the Tc2-like elements, that are conserved in Tc2 as well. This is also a midpoint rooted, bootstrapped parsimony tree. The same two clusters observed in figure 2.4 are still apparent in this tree. Tc2 clusters within one of these two groups. Table 2.8 contains a distance matrix for the Tc2-like elements. Between clusters elements are 70-77% identical. Within one cluster (containing f56d5) they are 96-98% identical. Within the second cluster they are 88-97% identical.

Analysis of Tc3 and Tc3-like sequences in the *C. elegans* genome:

Tc3 is 2335 bp long with 471 bp IR. Relevant features of Tc3 and related elements are summarized in Table 2.8. The top ten BLAST hits to Tc3 include:

Three elements similar in size to Tc3 with 467 bp IRs.

Two elements ~200bp shorter than Tc3 with 471 and 473 bp IRs.

Two elements 1368bp and 1360bp with 577 and 576 bp IRs respectively.

One 1827bp element with 477 bp IRs.

One element that was truncated by cosmid cloning that contains only the right IR of Tc3.

One element 1773 bp long with 479 bp IRs.

The IRs within each element are nearly perfect for most elements.

Table 2.9: Comparison of the length of IR, variation among the two IR of each element, and total length for Tc3 and Tc3-like cosmid sequences.

cosmid	IR (bp)	variable sites between IR	indels (bp)	length (bp)
b0303	467	4	1	2336
t02g5	467	5	0	2337
r10h1	467	2	0	2337
zk1086	-	-	-	732
t25g12	473	8	1	2166
zc64	471	2	1,2	2119
c25g4	477	6	1,1,6	1827
f27e5	577	4	0	1368
t13a10	576	10	4,7	1360
k10b3	479	2	12,1	1773
Tc3	471	3	0	2335

Appendix D contains an alignment of all Tc3 and Tc3-like sequences. The sequences very clearly fall into three separate groups. Gaps clearly distinguish the groups. There is obvious similarity between all elements in portions of the alignment. All three groups are similar over the first and last 200 bp of the alignment.

There are two groups of longer elements. One consists of sequences t25g12, zc64, c25g4 which appear to have a large segment of sequence that is similar to internal regions of Tc3 (coding region). The second group consists of full length and deleted versions of Tc3. Full length elements include b0303, r10h1, and t02g5. zk1086 is truncated, the Tc3 like sequence is contained at the very end of a cosmid and therefore this truncation represents a cloning artifact, not a real deletion at the end of the element. k10b3 looks like a Tc3 element with a large internal deletion.

There is one group of shorter elements containing sequences f27e5 and t13a10.

Within groups there is some length variation:

In the t25g12, zc64, c25g4 group:

c25g4 has a 6 bp insertion (in IR), 1bp insertion (in IR), a 351 bp deletion in the internal region and two 1 bp deletions (in IR) relative to the other elements of its type.

zc64 contains a 46 bp deletion in the internal region and a 2bp deletion in one IR relative to the others.

In the f27e5, t13a10 group there are a few small deletions:

f27e5 has a 3 bp internal deletion

t13a10 has 7 bp and 4 bp deletions in its left and right IRs respectively.

Among the Tc3, b0303, r10h1, t02g5, zk1086, k10b3 sequences there are a few length differences:

k10b3 has a 575 bp internal deletion.

zk1086 contains only the first 732 bp of Tc3 IR then the sequence is truncated (due to cosmid cloning).

t02g5 and r01h1 share a 1 bp insertion at position 315.

k10b3 has a 12 bp insertion in the right IR that consists of 12 Gs in a row.

Tc3 contains a unique 1 bp deletion in its right IR.

Figure 2.6 contains a midpoint rooted bootstrap tree for all of the Tc3 and Tc3-like elements. The tree was constructed using only the first and last 200 bp of the alignment. These sites are fairly similar among all of the elements. The tree obviously divides the sequences into the 3 groups already described. In addition there is some resolution within groups.

Table 2.10 shows a distance matrix for the sequences that appear to be Tc3 elements. The entire element sequences were considered in this analysis. The elements are all >99.6% identical to each other over their entire length (excluding gaps described above).

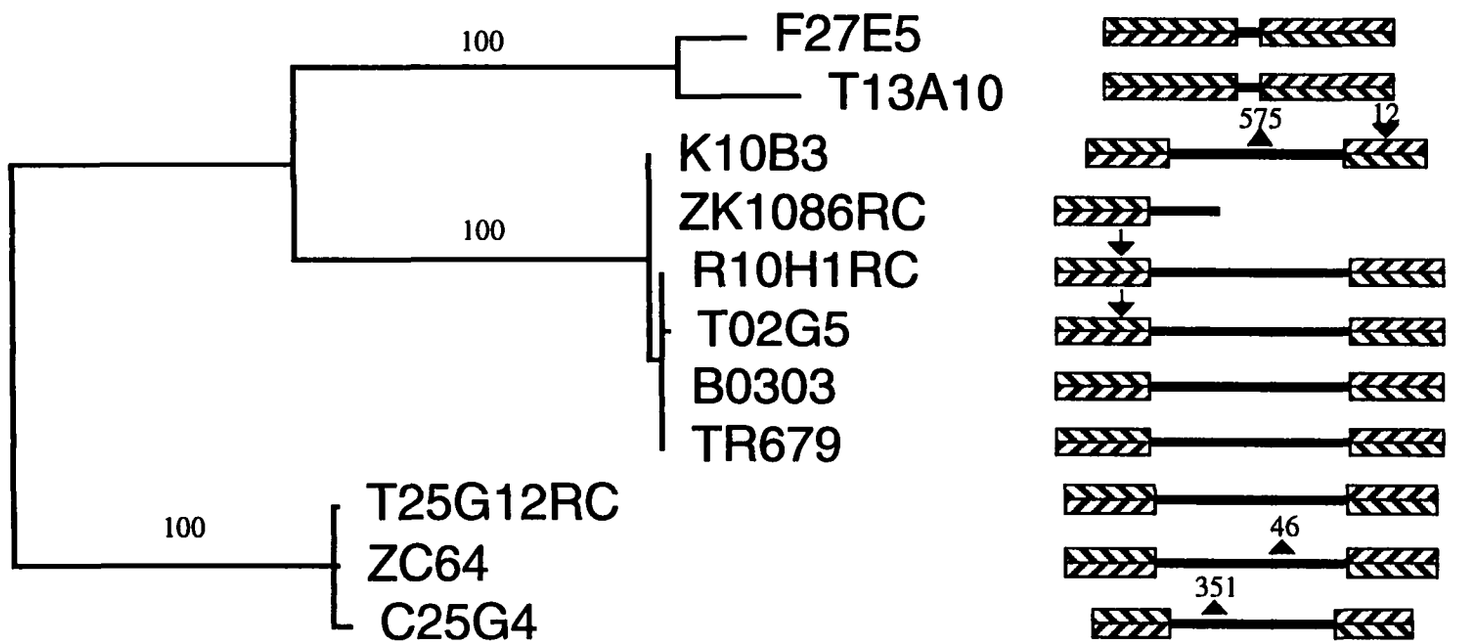


Figure 2.6 Parsimony bootstrap consensus tree (100 replicates) of Tc3 and related elements based on positions 1-200 and 2176-2376 from the alignment shown in Appendix D. The diagrams to the right of the tree show some of the major structural features of each element. Striped regions indicate IRs and insertions and deletions are indicated by an arrowhead with a number above it indicating the length of the indel (arrows pointing up are deletions and arrows pointing down are insertions).

Table 2.10: Pairwise distances between Tc3 and the four cosmid sequences with greatest similarity to Tc3 from the APPENDIX D alignment. Absolute distance are shown in the lower diagonal. Mean distances (adjusted for missing data) are shown in the upper diagonal.

	1	2	3	4	5	6
1 K10B3	-	0.000	0.002	0.002	0.003	0.002
2 ZK1086RC	0	-	0.005	0.004	0.005	0.003
3 R10H1RC	4	4	-	0.002	0.002	0.003
4 T02G5	3	3	4	-	0.002	0.003
5 B0303	5	4	4	4	-	0.003
6 Tc3	4	2	6	6	8	-

Table 2.11 is a distance matrix for the Tc3-like elements that contain an ORF. The t25g12 and zc64 elements are more similar to each other (99.5% identical) than they are to c25g4 (~98.5% identical to the other two elements), the element with a 351 bp internal deletion.

Table 2.11: Pairwise distances between three sequences from the APPENDIX D alignment that are shorter than Tc3 and encode a predicted protein that is similar to the Tc3 transposase. Absolute distance are shown in the lower diagonal. Mean distances (adjusted for missing data) are shown in the upper diagonal.

	1	2	3
1 T25G12RC	-	0.005	0.018
2 ZC64	10	-	0.015
3 C25G4	32	27	-

The two short elements, f27e5 and t13a10 are 89.2% identical over their entire length. They have almost identical structures with many single base changes scattered throughout the elements. Most of the element is IR, and most of the changes are in the IR (121 out of 148 differences are in the IRs). Both elements have the structure of foldback elements where f27e5 has 577 bp IRs with 214 bp internal sequence and t13a10 has 576 bp IRs with 208 bp internal sequence. In spite of the differences between the two elements, within each of these two elements, the IRs are nearly identical.

Tc3 has two ORFs. ORF1 is found at positions 727-1143 in the alignment followed by a small intron from 1144-1191 and ORF2 at 1192-1764. One of the groups of Tc3-like elements contains 2 similar ORFs with the intron in a conserved location. Table 2.12 shows differences between Tc3 transposases. There are no more than 5 variable sites. zk1086 is truncated after the first 5 codons of the first ORF.

Table 2.12: Variable sites from an alignment of predicted transposases for Tc3 and three Tc3-like elements. Each column heading indicates the ORF (1 or 2) followed by the position of the amino acid residue in that ORF. Sequences that match Tc3 are indicated by a quotation mark.

element	1.41	2.57	2.58	2.86	2.178
Tc3	V	L	L	N	F
R10H1RC	"	F	V	D	"
T02G5	E	F	V	D	"
B0303	E	F	V	D	I

Table 2.13 shows the distance matrix for the polypeptides encoded by the Tc3 elements and the Tc3-like element t25g12. The first ORF encodes 139 aa's, 60 of which vary between Tc3 and t25g12. 29 out of the 60 differences are within the first 65 aa's of the

Table 2.13: Pairwise distances between Tc3 transposase and the predicted amino acid sequence from four shorter cosmid sequences that also encode a signifcant ORF. Absolute distance are shown in the lower diagonal. Mean distances (adjusted for missing data) are shown in the upper diagonal.

	1	2	3	4	5
1 R10H1RC	-	0.003	0.006	0.009	0.391
2 T02G5	1	-	0.003	0.012	0.391
3 B0303	2	1	-	0.015	0.394
4 Tc3	3	4	5	-	0.391
5 T25G12RC	129	129	130	129	-

polypeptide which is known to contain a sequence specific DNA binding domain of Tc3 transposase. The second ORF encodes 190 aa's 70 of which vary between Tc3 and t25g12. k10b3 has a large deletion including part of ORF1, the entire intron, and some of ORF2. It is out of frame after ~55 amino acids and likely represents a transposase pseudogene. t25g12 has a single base change that alters the stop codon relative to the Tc3 elements. t25g12 encodes 15 extra C-terminal amino acids. c25g4 has a 351bp internal deletion including part of ORF1, the entire intron, and some of ORF2, and could encode a truncated polypeptide (first 99 amino acids of transposase). zc64 has a missense mutation at position 1538 (UGG trp -> UAG stop). It could produce a polypeptide more similar to t25g12 than Tc3. The polypeptide has a 75 amino acid C-terminal truncation relative to t25g12.

As noted by van Luenen et al. (1994) Tc3 contains a directly repeated sequence within the IRs. The first 29 bases of Tc3 match at 26 out of 29 positions with bases 176-202. In DNAse I footprinting experiments using the N-terminus of the Tc3 transposase it is sequences within these two regions of the IR that are protected. It is interesting to note that this appears to be a conserved motif across all of the Tc3-like elements. This suggests that similar transposases are acting on these elements (if they are even mobile).

Analysis of Tc4 and Tc4-like sequences in the *C. elegans* genome:

I examined the top ten BLAST hits to Tc4. f32a6 and f23g4 contain sequences with similarity to Tc4 but they lie at the end of a cosmid sequence and are incomplete, they will not be considered further. c27d11 and f36d4 sequences are similar to Tc4 except that in each case one IR seems to be deleted. Since they contain little if any internal sequences they are difficult to align to Tc4 and were not considered further in this analysis.

Tc4 is 1605 bp long with 775 bp IRs. Relevant features of Tc4 and related elements are summarized in Table 2.14. The remaining six BLAST hits include:

r04b3 an element with 368 bp IRs, 1476 bp long. Its IRs look like they could be longer (up to 773 bp) except that the left IR has a 138 bp deletion relative to the right.

f57g12 is 1400 bp long with 523 bp IRs. However, one end of the element has an additional 40 bp of sequence with strong similarity to the ends of Tc4. Hence, it looks as though the terminal 40 bp of IR is deleted from one end.

f49e11 is 1311 bp long with 473 bp nearly perfect IRs.

There are three short sequences. f23c11 is 895 bp long with 409 bp nearly perfect IRs, zk686 is 888 bp long with 403 bp nearly perfect IRs, t08g2 is 820 bp long with 137 bp IRs.

Table 2.14: Comparison of the length of IR, variation among the two IR of each element, and total length for Tc4 and Tc4-like cosmid sequences.

cosmid	IR (bp)	variable sites between IR	indels (bp)	length (bp)
Tc4	775	2	1,1	1605
f23c11	409	5	1	895
zk686	403	3	0	888
f49e11rc	473	1	0	1311
f57g12	523	7	40	1400
R04b3rc	773	20	138,1	1476
t08g2	137	2	2	820

Appendix E contains the alignment of these six Tc4-like elements with Tc4. Two of the shorter elements, f23e11 and zk686 are very similar over their entire length. t08g2, another short element, is most similar to Tc4 but contains several large deletions, (317 bp and 176 bp) in the left IR. There is a small island of similarity at position 439-460 that

breaks up the deletion into two peices. In addition, t08g2 contains a second large deletion (315 bp) in its right IR. The position of the ~315 bp deletions in the two IRs suggests that they are symmetrical deletions. One occurs 109 bp into left IR, the other, 111 bp into the right IR. The right IR has a 2 bp insertion relative to the left. t08g2 shows good alignment to Tc4 across the entire length of the element including portions within the IRs of the Tc4 element that are not within the IRs of t08g12.

f49e11 and f57g12 sequences look alike. They are similar over the entire length of the elements except for a 120 bp deletion in f49e11 at position 682 in the alignment and a 1 bp gap at position 339. f49e11 and f57g12 are very similar to Tc4 from positions 13-370 in the alignment and also from positions 781-1676. Both of these elements share deletions relative to Tc4. Their sequences are more similar to each other than to Tc4.

r04b3 looks like f49e11 and f57g12 in the region from 13-531 but from 532-780 it looks a lot more like Tc4 than f49e11 and f57g12. All of the sequences are alike from 780 to 1522. From 1523-1676 r04b3 looks more like f49e11 and f57g12.

Figure 2.7 is a tree showing the relationships among full length Tc4 and Tc4-like elements. The sequences form 2 distinct clades.

Table 2.15 contains a distance matrix for Tc4 and related sequences. f23e11 and zk686 are 99.4% identical. Tc4 and t08g2 are 93.4% identical. f49e11 and f57g12 are 96.2% identical.

Table 2.15 Pairwise distances between Tc4 and the six Tc4-like cosmid sequences from the APPENDIX E alignment. Absolute distance are shown in the lower diagonal. Mean distances (adjusted for missing data) are shown in the upper diagonal.

	1	2	3	4	5	6	7
1 F23c11	-	0.006	0.373	0.388	0.364	0.228	0.242
2 Zk686	5	-	0.370	0.385	0.361	0.224	0.241
3 F49e11rc	301	299	-	0.038	0.066	0.183	0.151
4 F57g12	333	331	50	-	0.116	0.223	0.187
5 R04b3rc	315	313	87	165	-	0.172	0.112
6 T08g2	126	124	124	177	138	-	0.067
7 Tc4	212	211	193	261	161	54	-

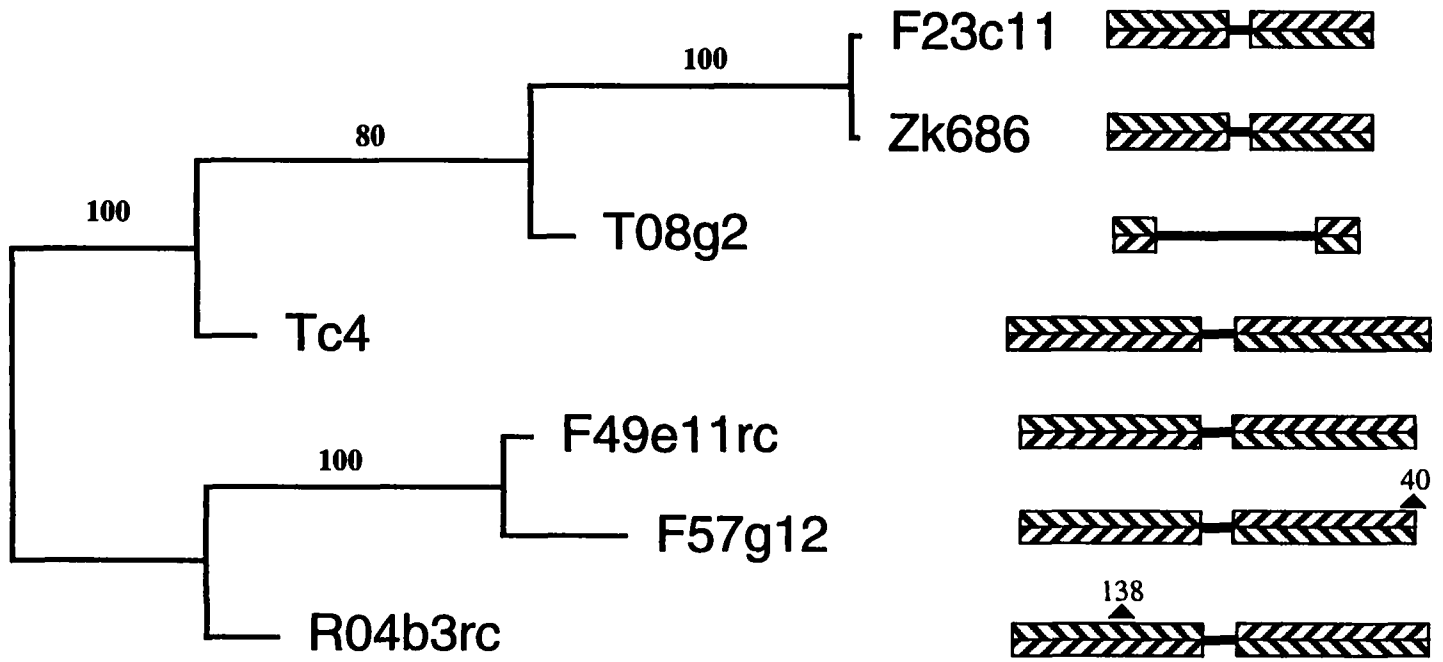


Figure 2.7 Parsimony bootstrap consensus tree (100 replicates) of Tc4 and related elements based on positions 14-1666 from the alignment shown in Appendix E. The diagrams to the right of the tree show some of the major structural features of each element. Striped regions indicate IRs and insertions and deletions are indicated by an arrowhead with a number above it indicating the length of the indel (arrows pointing up are deletions and arrows pointing down are insertions).

Analysis of Tc5 and Tc5-like sequences in the *C. elegans* genome:

Tc5 is 3171 bp long with 491 bp perfect terminal IRs. Relevant features of Tc5 and related elements are summarized in Table 2.16. The next ten best BLAST hits include:

t13c2, a 3193 bp long element with 435 bp nearly perfect IRs.

four elements t19d7, t14g8, c01b7, and c48b4 ranging in size from 1423-1632 bp long with near perfect terminal IRs of 666-770 bp.

five small elements c04e7, c24a3, f44b9, zk930, and c39d10, 556-681bp long with near perfect 99-127 bp terminal IRs.

Table 2.16: Comparison of the length of IR, variation among the two IR of each element, and total length for Tc5 and Tc5-like cosmid sequences.

cosmid	IR (bp)	variable sites between IR	indels (bp)	length (bp)
Tc5	491	0	0	3171
T13c2	435	0	1	3193
C04e7	127	10	0	632
c24a3	111	7	0	681
f44b9	99	6	0	627
zk930	101	2	0	592
c39d10	120	7	0	556
t19d7	770	5	0	1632
t14g8	758	3	2	1606
c01b7	757	16	1	1607
c48b4	666	2	1	1423

I aligned all 11 sequences (not shown) together (using copies of Tc5 and t13c2 sequences with large deletions in the middle of the elements to reduce difficulties in

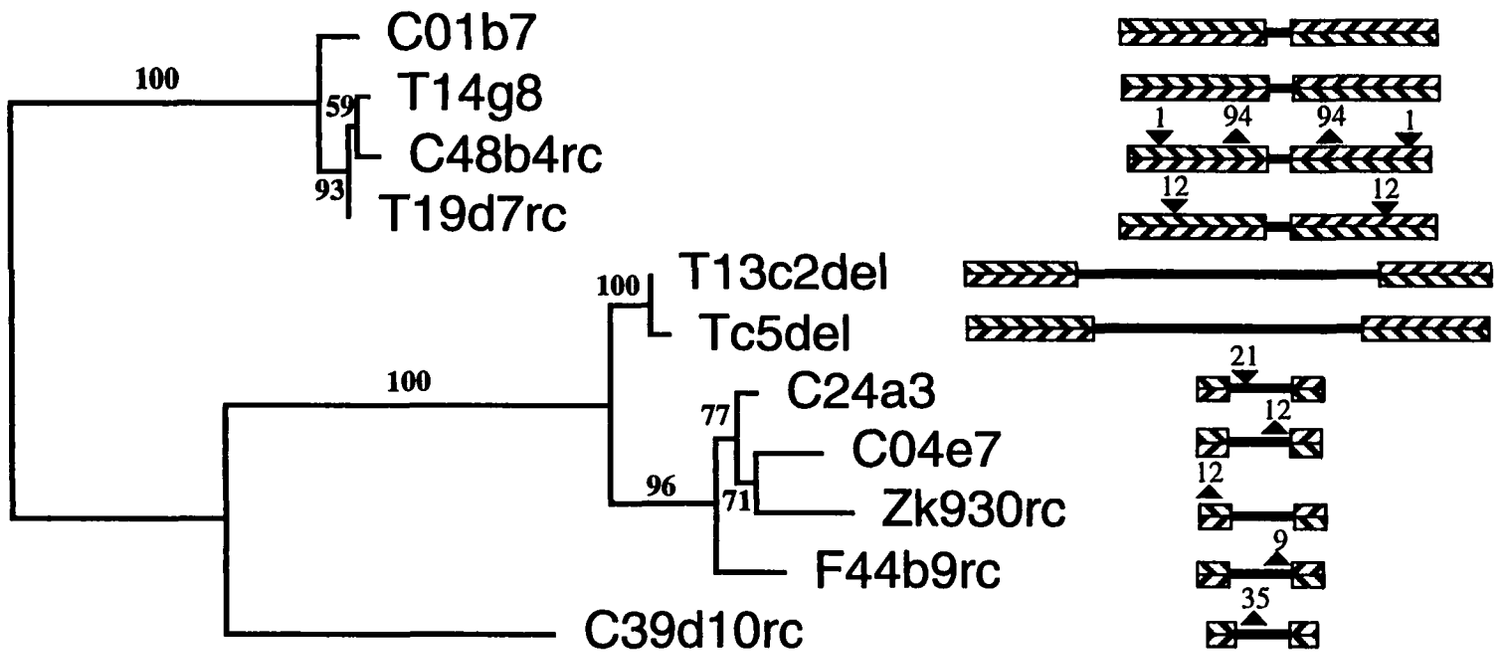


Figure 2.8 Parsimony bootstrap consensus tree (100 replicates) of Tc5del and related elements. The diagrams to the right of the tree show some of the major structural features of each element. Striped regions indicate IRs and insertions and deletions are indicated by an arrowhead with a number above it indicating the length of the indel (arrows pointing up are deletions and arrows pointing down are insertions). Note that the drawings of Tc5 and T13c2 are of the full length elements.

aligning sequences). The sequences clearly fall into three groups. One group consists of the two long elements, Tc5 and t13c2. A second group consists of small elements that match approximately 135 bp at each end of Tc5 but contain no significant matches to internal regions of the larger elements. The third group consists of larger elements that match over the entire IRs of Tc5 (491 bp). These elements have IRs longer than Tc5, with their internal regions showing no obvious similarity to sequences in Tc5 or the smaller elements.

Figure 2.8 contains a tree showing the relationship among all of the Tc5-like sequences. The tree is interesting because it groups the Tc5 elements (Tc5 and t13c2) with the short elements that have small IRs to the exclusion of the larger fold-back like elements.

Appendix F contains a pairwise alignment of Tc5 and the element contained on cosmid t13c2. There are 50 nucleotide differences most of which are clustered between positions 431-513 in Tc5, the same region that contains three small insertions (7 bp,8 bp,6 bp) in t13c2 relative to Tc5. There are two single base deletions at positions 2728, 2853 in Tc5. At position 612 in Tc5 there is a 2 bp insertion in t13c2. The 491 bp right IR of Tc5 is almost identical to 491 bp at the right end of t13c2. The IR of t13c2 were described as 435 bp owing to a deletion after position 435 in the t13c2 sequence relative to Tc5.

To examine if the changes between the two Tc5 elements affects their transposase coding sequence I compared the amino acid sequences. Tc5 encodes a predicted 532 aa polypeptide whereas t13c2 is predicted to encode 728 aa polypeptide with the size difference occurring at the C-terminus. The only other differences are 3 aa replacements: Q144R, M308K, and L365Q. Changes are shown as Tc5->t13c2.

Appendix G contains an alignment of 4 Tc5-like elements with long IRs and a foldback structure. There are several interesting gaps in the sequences of these elements shown in Table 2.17. The 1 bp insertions at positions 237 and 1419 in c48b4 as well as the 94 bp deletions at positions 433 and 1129 are within the IRs of this element and are symmetrical.

Table 2.17: Describes the position of insertions and deletions among Tc5 related elements from the alignment in Appendix G. The indels marked with a * represent symmetrical insertions and deletions within an element.

position in alignment	indel	contained in element
237*	+1	c48b4
272*	+12	t19d7
433*	-94	c48b4
973	+1	c48b4
1022	-2	t14g8
1129*	-94	c48b4
1371*	+12	t19d7
1419*	+1	c48b4
1625	-1	c01b7

Likewise, in t19d7 the 12 bp insertions at positions 272 and 1371 are symmetrical. These Tc5-like elements have a foldback structure, with long IRs, and all have an internal non-IR segment of 91-93 bp.

Table 2.18 shows a distance matrix for these four related elements. Sequences are all 98.1% to 99.2% identical, excluding gaps, over their entire length.

Table 2.18: Pairwise distances between Tc5-like cosmid sequences for positions 17-1653 of the APPENDIX G alignment. Absolute distance are shown in the lower diagonal. Mean distances (adjusted for missing data) are shown in the upper diagonal.

	1	2	3	4
1 C01b7	-	0.013	0.019	0.016
2 T14g8	21	-	0.014	0.008
3 T19d7rc	31	22	-	0.013
4 C48b4rc	22	12	19	-

Figure 2.9 contains a boostrap tree of the four Tc5-like elements with long IRs. Most of the variable sites are different in only one element, a similar situation to the gaps, so there are very few informative sites in the alignment. The tree is a polytomy with long

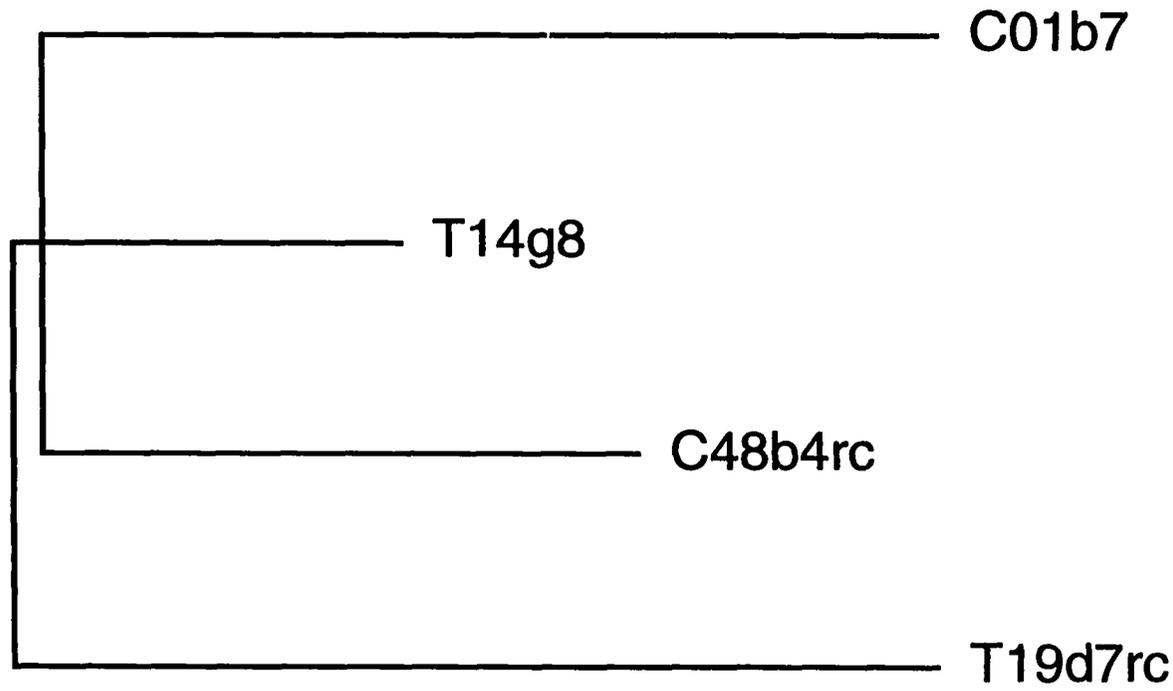


Figure 2.9 Parsimony bootstrap consensus tree (100 replicates) of Tc5-like foldback elements based on positions17-1673 from the alignment shown in Appendix G.

terminal branches and no significant bootstrap values.

Appendix H contains an alignment of shorter Tc5-like elements. The alignment reveals some size variation between sequences. The positions of insertions and deletions among these elements are shown in Table 2.19. None of these indels appear to be symmetrical like the ones seen among the other group of Tc5 related elements.

Table 2.19: Describes the position of insertions and deletions among Tc5 related elements from the alignment in Appendix H. Note that indels are not shown with respect to any particular reference sequence.

position in alignment	indel	contained in elements
26	-12	zk930
158	-48	c39d10
161	-33	f44b9, zk930
173	-29	c04e7
262	-1	c04e7
281	-35	c39d10
291	-2	c24a3, f44b9
296	-4	c24a3, f44b9, c04e7
335	+1	c04e7
359	-19	c39d10
413	-6	c24a3
413	-5	f44b9, c04e7
413	-2	zk930
475	-1	c39d10
487	+1	f44b9, c39d10
519	-3	zk930
520	-12	c04e7
521	-2	f44b9
541	+12	c39d10
606	-19	f44b9
615	+10	c24a3

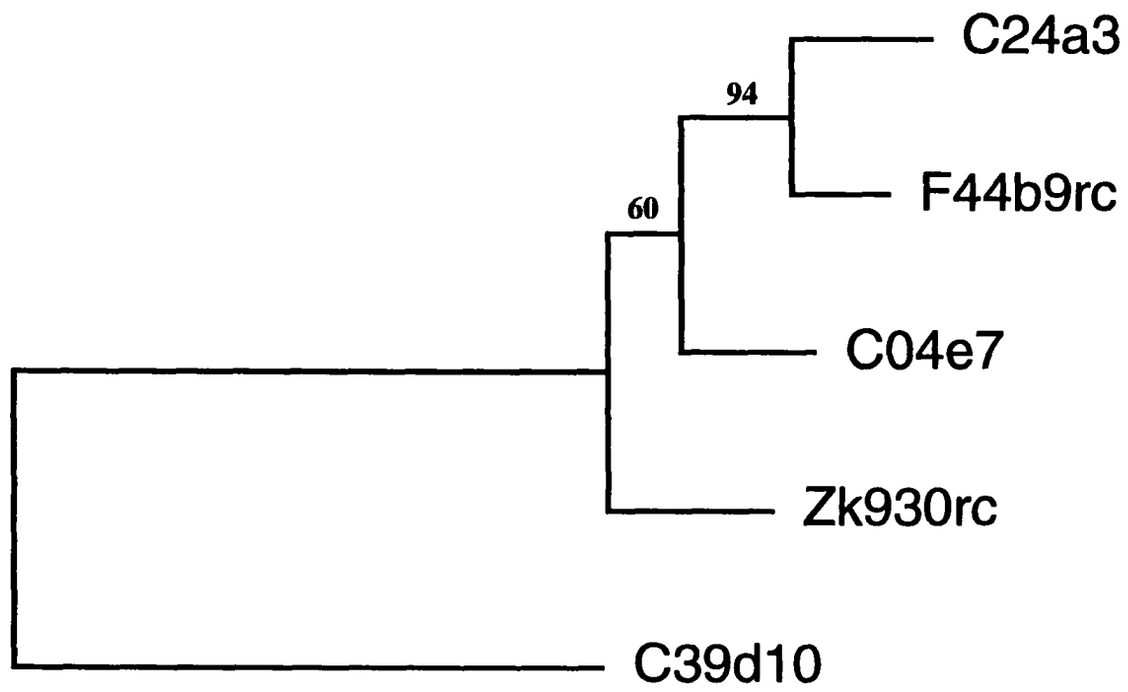


Figure 2.10 Parsimony bootstrap consensus tree (100 replicates) of short Tc5-like elements based on positions 12-717 of the alignment shown in Appendix H.

Table 2.20 contains a distance matrix for short Tc5-like elements. c24a3, f44b9, and c04e7 are all ~93% identical. zk930 is a bit more divergent owing to a deletion in one terminus (the alignment may include cosmid sequence that is not part of this element) and c39d10 is ~60% identical to all the other sequences.

Table 2.20: Pairwise distances between five short Tc5-like cosmid sequences for positions 12-717 of the APPENDIX H alignment. Absolute distance are shown in the lower diagonal. Mean distances (adjusted for missing data) are shown in the upper diagonal.

	1	2	3	4	5
1 C24a3	-	0.062	0.089	0.117	0.375
2 F44b9rc	39	-	0.074	0.096	0.363
3 C04e7	56	45	-	0.094	0.355
4 Zk930rc	73	59	57	-	0.374
5 C39d10	215	205	199	210	-

Figure 2.10 shows a tree constructed using the entire short Tc5-like element sequences. c24a3 and f44b9 reliably cluster to the exclusion of the other elements. There is also some support for a clade that includes c04e7.

Analysis of Tc6 and Tc6-like sequences in the *C. elegans* genome:

Table 2.21 contains relevant features of Tc6 and related elements identified as the top 9 highest scoring BLAST hits. They include:

Five elements are almost the same length as Tc6. They range from 1591-1605bp long and contain near perfect IRs ranging from 740-766bp long. f53b7 has slightly more degenerate IRs than the other sequences. It contains 25 changes in nucleotide sequence and three small indels between its IRs.

One element that is 1048bp with perfect 421bp IRs.

Three small elements 424-954bp long which show similarity to only one IR of Tc6.

Appendix I contains an alignment of Tc6 and Tc6-like elements. A large number of insertions and deletions are observed between different copies of these elements. Some of

Table 2.21: Comparison of the length of IR, variation among the two IR of each element, and total length for Tc6.1 and Tc6-like cosmid sequences.

cosmid	IR (bp)	variable sites between IR	indels (bp)	length (bp)
Tc6.1	766	1	0	1603
zk669	766	3	0	1603
zk180	766	2	1	1598
zc395	421	0	0	1048
f53b7	740	25	1,5,9	1591
w03a3	758	8	1	1593
f48e8	764	8	2,1	1605
c33h5	-	-	-	848
ac3	-	-	-	954
t26a8	-	-	-	424

the changes are unique to a particular sequence whereas others are shared between different sequences. Table 2.22 shows positions in the alignment which contain gaps. Note that in the alignment at t26a8 ends position 439, c33h5 ends at position 870, and Ac3 ends at position 1260. Gaps in the region 230-940 in Ac3 were ignored since the sequence aligns very poorly to the others in this region despite good similarity at both ends of Ac3

Table 2.23 contains a distance marix for Tc6 and related elements. Tc6, zk669, and zk180 differ from each other at a maximum of 4 sites over the entire alignment (ignoring gaps). Among all of the "full length elements" the maximum difference is 132 out of 1591 bp (91.7% identical). Ac3 is clearly the most divergent sequence showing ~65% identity to the full length elements over the entire alignment.

Figure 2.11 contains a tree of Tc6 and the related elements constructed from sites that appear conserved among all sequences. This conserved region is from position 12-225 in the alignment. The tree contains two clusters one containing the full length elements and c33h5 and a second cluster with t26a8 and zc395.

Figure 2.12 contains a tree constructed using the full length Tc6 elements. This tree gives better resolution within the groups. One cluster contains three almost identical Tc6 elements, w03a3 is the next most similar to these three.

All sequences have the structure of foldback elements with IRs ranging from 740-766 with internal regions of 71-111. zc395 has 421 bp IRs and 206 bp internal sequence because it appears to have a deletion that makes its IRs shorter and its internal region longer relative to other elements.

Table 2.22: Describes insertions and deletions among Tc6 related elements from the alignment contained in Appendix I. Note that no particular sequence is used as a reference for determination of indels.

position in alignment	indel	contained in elements
160	-1	c33h5, f48e8, f55b7
165	+2	Ac3
177	-1	t26a8
198	-1	f48e8
199	-1	w03a3
316	+1	t26a8
438	-556	zc395
461	-5	w03a3
546	+2	c33h5
610	+1	c33h5, f48e8, f55b7
669	-2	c33h5, f48e8
675	-1	f55b7
723	+1	c33h5, f48e8, f55b7
763	-1	f55b7
787	+1	f55b7
805	+2	c33h5
863	-5	zk180
920	+1	f48e8, w03a3, f55b7
965	-2	f48e8
1028	-1	Tc6, zk669, zk180, w03a3
1072	-9	f55b7
1172	-5	w03a3
1176	-5	f55b7
1233	+1	f48e8
1381	-1	zk180
1436	-1	w03a3
1442	+1	f48e8
1474	-1	f48e8

Table 2.23: Pairwise distances between Tc6.1 and Tc6-like cosmid sequences for positions 12-1627 of the APPENDIX I alignment. Absolute distance are shown in the lower diagonal. Mean distances (adjusted for missing data) are shown in the upper diagonal.

	1	2	3	4	5	6	7	8
1 T26a8	-	0.028	0.038	0.048	0.066	0.069	0.066	0.057
2 Zc395	12	-	0.029	0.030	0.066	0.068	0.067	0.063
3 C33h5	16	12	-	0.027	0.072	0.073	0.070	0.075
4 F48e8	20	31	23	-	0.068	0.070	0.070	0.074
5 Tc61	28	70	61	109	-	0.001	0.001	0.021
6 Zk669rc	29	72	62	111	2	-	0.003	0.022
7 Zk180rc	28	71	59	111	2	4	-	0.021
8 W03a3	24	66	63	117	33	35	34	-
9 F55b7	30	80	71	132	86	88	87	81
10 Ac3	161	211	267	328	323	324	322	320

	9	10
1 T26a8	0.071	0.382
2 Zc395	0.077	0.304
3 C33h5	0.084	0.423
4 F48e8	0.083	0.343
5 Tc61	0.054	0.336
6 Zk669rc	0.055	0.338
7 Zk180rc	0.055	0.335
8 W03a3	0.051	0.337
9 F55b7	-	0.340

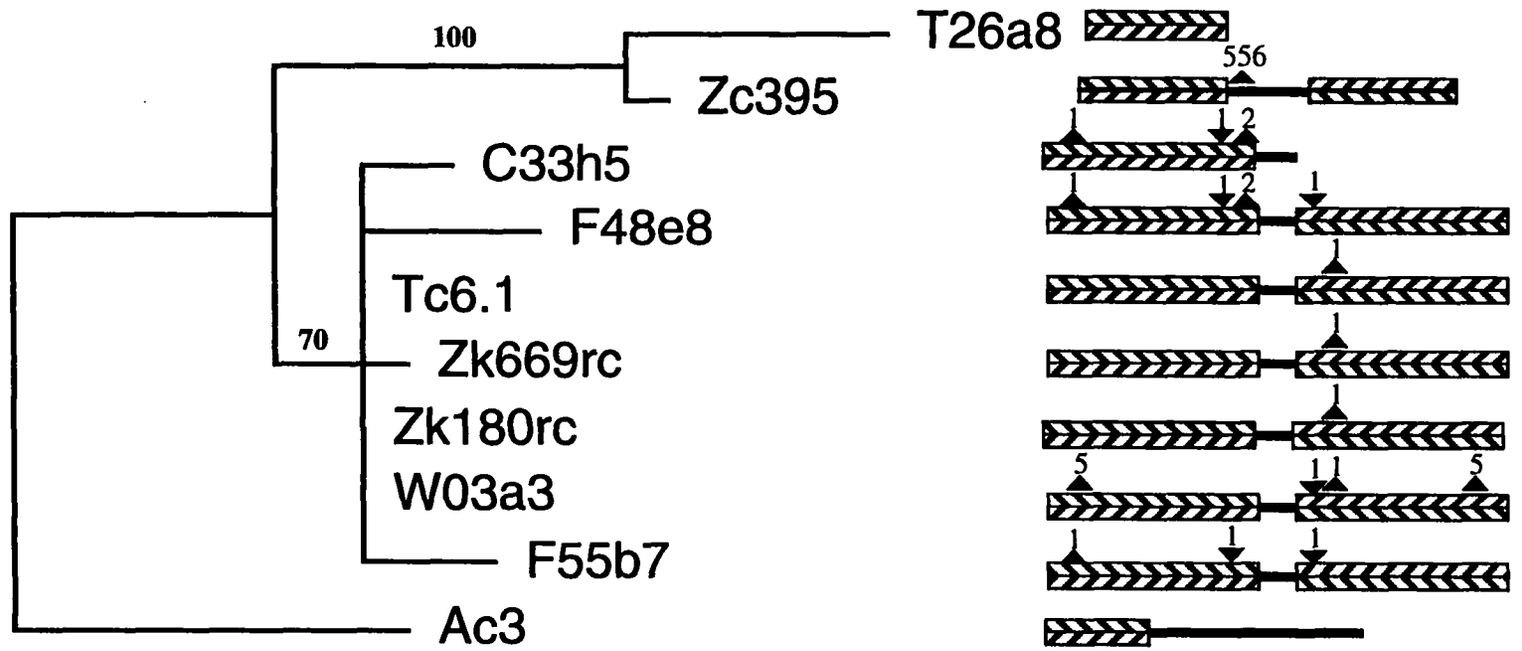


Figure 2.11 Parsimony bootstrap consensus tree (100 replicates) of Tc6 and related elements based on positions 12-225 of the alignment shown in Appendix I. The diagrams to the right of the tree show some of the major structural features of each element. Striped regions indicate IRs and insertions and deletions are indicated by an arrowhead with a number above it indicating the length of the indel (arrows pointing up are deletions and arrows pointing down are insertions).

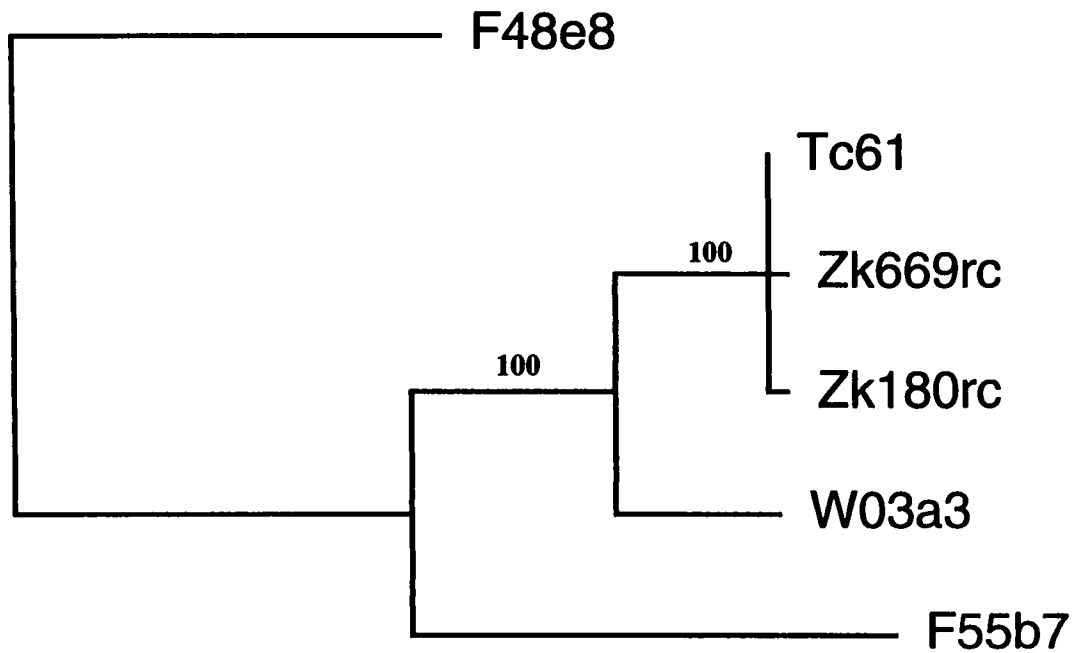


Figure 2.12 Parsimony bootstrap consensus tree (100 replicates) of Tc6 and related elements based on positions 12-1667 from the alignment shown in Appendix I.

Conclusions:

The *C. elegans* genome is replete with transposons. There is a surprising amount of sequence variation among different copies of a transposon, and in fact, of the 60 or so sequences considered in this analysis, no two were identical over their entire length. Among the different families of transposable elements there seem to be groups of autonomous elements, in this case defined as elements capable of encoding a transposase, as well as nonautonomous elements that contain termini identical to the ends of an autonomous element but do not contain coding sequence. Extensive genetic analysis will be required to determine if the elements considered in this analysis share these sort of relationships. In some cases the autonomous element associated with the nonautonomous element has not been identified or does not exist.

The group of Tc1 and related sequences appears to include an autonomous element, Tc1, and a nonautonomous element with 38bp terminal IRs like Tc1's and a fold-back structure. In addition, the sequence on cosmid c30g4 may represent a degenerate fold-back element. The Tc1 elements and their associated foldback elements are among the most highly conserved of the elements considered. Even so, there is variation between copies, even within the coding region of Tc1.

The Tc2 element sequenced by Ruvolo et al. (1992) remains as the only example of what is likely to be the autonomous element related to the nonautonomous Tc2 elements described in this chapter. The Tc2 nonautonomous elements cluster into two groups, one of which contains Tc2. This suggests that the two classes of Tc2 nonautonomous elements may have independent origins, possibly from different copies of an autonomous element.

The Tc3 related sequences are unique in containing what appear to be two distinct autonomous elements, Tc3 as well as a slightly smaller element that encodes a similar transposase. There also appears to be a nonautonomous Tc3 like element with a fold-back

structure. The Tc3 elements all contain two conserved blocks of sequence that are likely to play a role in transposase binding to element sequences. The structure of this region is conserved but the nucleotide sequence in this region varies somewhat between elements. This suggests that the elements are recognized by different transposases, but may interact with the transposase in a similar manner.

None of the Tc4-like elements considered in this study appear to be autonomous. The elements considered here all appear to be members of families of Tc4-like nonautonomous elements. Although not considered in these analyses, an element has been described in the *C. elegans* genome that has the expected features of an autonomous Tc4 element. Li and Shaw (1993) characterized a variant Tc4 element designated Tc4v. Tc4v has IRs similar to Tc4 however, disrupting one of these IRs is a long ORF capable of encoding a polypeptide that shares significant similarity to the product encoded by Tc5. The Bristol genome contains several copies of Tc4v, but none of them were contained in the subset of the genome used in these analyses.

The analysis of sequences related to Tc5 revealed a slightly divergent copy of Tc5, a presumed autonomous element, as well as what appear to be two families of nonautonomous Tc5-like elements with very different structures. There appears to be one group of fold-back elements and a second group of small Tc5-like elements with shorter IRs. The two Tc5 elements encode very similar transposases except that the polypeptide encoded by t13c2 is longer than the product predicted for Tc5.

Tc6 is a fold-back element and is presumably non-autonomous. Thus far, no putative autonomous elements with similarity to Tc6 have been identified in the genome, and in fact no direct evidence for Tc6 transposition exists. There may be no element in the genome capable of directing Tc6 movement. This could explain the high levels of sequence diversity detected among Tc6-like sequences in the genome. Tc6 elements may represent the vestiges of a once active transposon family. Loss of the autonomous copy of an

element may render nonautonomous elements incapable of movement and subject to decay by a steady accumulation of mutations.

The idea that these putative nonautonomous elements are inserted using the same factors that control the autonomous elements is strengthened by the observation that within a group of related elements, all members, whether or not they contain an ORF, insert into the same target site. Tc1, Tc2, Tc3, Tc6 and all of the sequences related to these elements insert into a TA and appear to duplicate those bases upon insertion. Tc4, Tc5, and their related elements all insert into the sequence TNA and appear to duplicate this target sequence upon insertion.

No evidence for detectable levels of transposon activity exists for the Bristol strain. However, many of the element families found in this genome actively transpose in other strains. The reasons for the differences in transposon activity between strains is unknown. According to my analysis, none of the transposons examined in the Bristol genome have a sequence identical to the sequence of an element identified as a new insertion in another strain. Therefore, it is possible that the differences in activity between strains is due to differences in sequence between elements in different genomes. However, it is known that elements in the Bristol genome actively excise in somatic tissues, and can actively transpose in the soma when transposase is overexpressed suggesting that these elements contain the cis-elements necessary for activity. In addition, my analysis suggests that many of the Bristol transposons contain ORFs that could encode full length transposases. Therefore it is possible that the differences in activity between strains are due to changes in host-encoded factors that regulate transposon expression or activity and not changes in the elements themselves.

The sequences considered in these analyses form groups of related elements, often with very different structures. In particular, elements, such as Tc1, a transposon with short IRs (54 bp), seem to be related to elements with much larger IRs (e.g. the elements in Table 2.2

with 348 bp IRs). If these elements share a common origin, it suggests that IR sequences have expanded or contracted giving rise to the observed elements. Comparisons among related elements in these analyses reveal several mechanisms that could be responsible for changing IR structures. In some cases, indels were observed in one member of a pair of inverted repeats (e.g. r04b3, see figure 2.7). This process can lead to a shortening of IRs within an element since one IR has a region that no longer pairs with the other. Symmetrical insertions and deletions observed in some elements (e.g. in c48b4 and t19d7, see figure 2.8) suggest another mechanism involved in IR evolution. Chance occurrence of indels in corresponding regions of the two IRs of an element seems unlikely. A more likely explanation for symmetrical indels in these elements is mismatch repair of IR sequences when paired. If one IR contains an indel with respect to the other, repair of the mismatch during pairing of IRs could give rise to symmetrical insertions or deletions depending on which IR is used as a template for repair.

This chapter serves as a preliminary investigation of the relationships among transpsons in the *C. elegans* geneome. When the genome sequence is complete, analyses similar to those presented in this chapter will be extremely useful in reconstructing the relationships among the transposon sequences discovered. However, establishing times of divergence between element sequences requires information that is unlikely to emerge from the sequence of a single nematode genome. Ideally transposon sequences from other *C. elegans* strains and closely related nematodes could be compared with the Bristol sequences for more complete phylogenetic resolution.

CHAPTER III

ATTEMPTS TO CHARACTERIZE THE PHENOTYPIC CONSEQUENCES OF TRANSPOSABLE ELEMENT INSERTION

Summary:

This chapter describes a set of experiments designed to address the phenotypic consequences of element insertion. The goal of the experiment was to isolate a large number of independent germ-line insertions into a set of *C. elegans* genes and ascertain their phenotypic effect. Ultimately the method chosen to isolate insertions, sib-selection PCR, was found to be impractical for collection of a large number of insertions. Screens for new element insertions required great effort, and resulted in a large proportion of false positives. Many insertions were detected, but attempts to isolate the animals containing the insertions were largely unsuccessful. The reason animals containing insertions are difficult to isolate is due to high levels of somatic Tc1 activity, which is the focus of CHAPTER IV.

Section 1 in this chapter will outline the rationale and objectives of the experiments. Section 2 will discuss the sib-selection PCR method used to identify and isolate new transposon insertions. Section 3 describes the results of these experiments and discusses the difficulties encountered as well as possible improvements for the sib-selection PCR technique.

Introduction:

Numerous studies in diverse taxa clearly demonstrate that transposable elements are a significant source of genetic variation. The precise nature of this genetic variation and its consequences for host and element evolution remains unclear. I wanted to address the

consequences of transposon insertion using *C. elegans* as a model system. Specifically, I wanted to know how often transposon insertions into coding regions of a gene result in a mutant phenotype, and why we observe the resulting phenotype (or lack of a phenotype).

Genetic methods may underestimate the level of transposon activity and the range of phenotypic variation elements can generate.

Current estimates of the rates of transposon insertion and excision in various organisms are based on measurements using genetic methods. These estimates rely on the largely untested assumption that most transposon insertions occurring in coding sequences lead to a disruption of gene function and that element excision usually results in genetic reversion. If element insertion and excision events lack phenotypic consequences, element activity may be considerably higher than predicted by genetic methods.

Two lines of evidence support this idea. First, Engels and co-workers (1990) demonstrated that P-element excision in *Drosophila melanogaster* is much more frequent than predicted by measures of phenotypic reversion. These studies revealed that most transposon excision events are silent. In animals homozygous for an element insertion, the repair process that heals the double strand break generated when an element excises uses the homologous chromosome (or possibly sister chromatid) as a template, usually restoring a copy of the element to the excision site in the process. Second, studies in maize, *Drosophila*, *C. elegans* and mice demonstrate that transposon insertions can function as introns; element sequences can be spliced from pre-mRNA (Kim et al., 1987; Steinmeyer et al., 1991; Kobayashi et al., 1993; Purugganan, 1993; Rushforth et al., 1993; Rushforth and Anderson, 1996), often yielding partially, and in some cases fully functional protein products. These studies suggest that many transposon insertions in exons have no phenotypic effect because splicing removes the insertion from transcripts.

Sib-selection/PCR can isolate insertions without regard to phenotype

To estimate the proportion of element insertions that disrupt coding sequences,but do not cause a mutant phenotype, I tried to isolate transposable element insertions into target genes for which the loss-of-function phenotype is well characterized. I wanted to isolate the insertions by virtue of the molecular structure of the resulting alleles, without regard for phenotype. For each new insertion allele I hoped to characterize the phenotypic consequences and determine the fate of element sequences in gene transcripts. A method developed recently in both *Drosophila* (Ballinger and Benzer, 1989; Kaiser and Goodwin, 1990) and *C. elegans* (Rushforth et al., 1993; Zwaal et al., 1993) provides a way to identify new transposon insertions without regard for a phenotype. This approach combines the genetic method of sib-selection with the polymerase chain reaction to identify transposon insertions in any gene for which some nucleotide sequence is known. Details of the procedure are described in the next section. For the present discussion, it is important to note that the inspiration for the development of this technique was to establish a method to determine the loss-of-function phenotype for any cloned gene. These approaches were based on the assumption that most or all transposon insertions into a gene will generate null mutations. Ironically, the results reported in one of these studies (Rushforth et al., 1993) provides additional evidence that this is not the case. Using a sib-selection PCR protocol, five insertions of the *C. elegans* transposon Tc1 were isolated in two different genes, three in *mlc-2* and two in *hlh-1*. All five insertions were in exons and in each case the resulting phenotype was wild-type. Further analysis of the *mlc-1*::Tc1 strains revealed that in each case Tc1 is spliced from *mlc-1*::Tc1 transcripts, leaving small in-frame insertions or deletions in the mRNA. These results are consistent with the hypothesis that transposon insertions in exons are often silent due to splicing of the insertion. This interpretation is strengthened by the recent demonstration that the loss-of-function phenotype for both of these genes is lethal (Rushforth and Anderson pers. comm).

As transposon-based gene disruption techniques are applied to more genes in these critical model organisms, and extended to other organisms, it will be important to understand the relationship between transposon insertion and mutant phenotype. The results described above reinforce the need for a systematic analysis of the question, using genes with convenient and well established null phenotypes.

Muscle genes are good targets

Genetic analysis of muscle function is difficult in many systems because mutations in muscle genes are often lethal or difficult to propagate. *C. elegans* has become a good model for genetic investigation of muscle function due in part to its mode of reproduction as a self-fertile hermaphrodite. Since worms do not have to be able to move in order to reproduce, even mutations resulting in severe paralysis can be propagated. Many mutations affecting muscle structure and function have been described, and several genes and proteins are well characterized. Two genes have been the focus of numerous studies. *unc-54* encodes a myosin heavy chain protein found in *C. elegans* body muscle and *unc-54* loss-of-function mutants are paralyzed, flaccid, and egg laying defective. *unc-22* encodes a protein, twitchin, thought to be involved in regulating muscle activity. *unc-22* loss-of-function mutants display a continuous fine twitching of body wall muscle. Both *unc-54* and *unc-22* have been cloned and sequenced. Because of the easily identified mutant phenotypes associated with *unc-54* and *unc-22* mutations, these genes have proved useful in studies of transposon activity. Several germ-line Tc1 insertions have been isolated in *unc-54* and *unc-22* by virtue of the mutant phenotype generated upon element insertion. Element excision from these genes has been examined by monitoring phenotypic reversion from transposon induced mutant phenotypes.

I chose to address the phenotypic consequences of element insertion into *unc-54* and *unc-22* because of their well characterized mutant phenotypes as well as the wealth of

information concerning transposon insertion and excision for these two loci. The fact that several Tc1 insertions into each of these genes result in a mutant phenotype indicates that at least some proportion of insertions in this gene will disrupt its function. I wanted to determine the proportion of insertions into these genes which lack a phenotypic effect. Using a technique that does not rely on a mutant phenotype to detect new insertions I hoped to compare the distribution of insertion sites to those observed when screening for insertions by phenotypic criteria.

Tc1 is active in the germline of *mut-2* animals

Tc1 activity is regulated in strain specific and tissue specific manner. This feature can be useful in the manipulation of transposon insertion alleles. Insertions are isolated in mutator strains where elements transpose in the germline. To stabilize the insertion allele (i.e. prevent its excision) the mutant strain can be backcrossed to a strain where the element is not active. Subsequent reactivation of insertion alleles can be accomplished by introduction of a mutator background. *mut-2* mutator strains exhibit the highest levels of germ-line transposition of Tc1. To increase the likelihood of observing new insertion events I used the *mut-2(r459)* mutator strain TW186.

Methods:

Two variations of the sib-selection PCR protocol (Rushforth et al., 1993; Zwaal et al., 1993) were used to try to isolate germ-line Tc1 insertions into the *unc-22* and *unc-54* loci. The first, and less successful, method involved PCR and Southern blotting to detect new insertion events. The second, slightly more successful, method used a nested PCR protocol to detect insertion events.

Both methodologies rely on the same basic principles. Gene specific and transposon specific primers are designed in such a way as to allow amplification only when a

transposon inserts into a gene of interest. PCR is performed on DNA from one half of a population of animals using gene and transposon specific primers to detect new insertion events. If an insertion is detected in half of the animals, the remaining half is subdivided, cultured, and again screened for the insertion. The process of screening and subdividing is repeated until an entire population of animals homozygous for the insertion is obtained.

Sib-selection PCR with Southern blotting to isolate Tc1 insertions in *unc-54*

unc-54 was chosen as the first target to isolate new Tc1 insertions. I hoped to use a set of primers covering most of the *unc-54* coding region to isolate new germ-line insertions of Tc1 into many sites in the gene. Positions of *unc-54* primers and Tc1 primers used in the PCR are shown in figure 3.1.

50 populations of TW186 *mut-2(r459)* animals were grown on 60mm petri dishes containing nematode growth media seeded with *E. coli* strain OP50. Each population was started with approximately 50 L3 larvae. Worms were grown until the bacterial lawn was cleared. At this point there are approximately 5000 animals, of mixed stages, on each plate. Worms were harvested from petri dishes in 1ml M9 medium. 0.33ml of the worm suspension was placed on a fresh seeded plate. The second 0.33ml were frozen; DNA was prepared from these samples only when a potential insertion was detected from a particular population. The remaining 0.33ml of worms in M9 was used for DNA preparation. Worms were centrifuged briefly, M9 was removed and the worm pellet was washed in 0.5ml M9 centrifuged, washed in 1ml water, centrifuged, resuspended in 1ml WLB, centrifuged, and resuspended in 200ul WLB. DNA preps were frozen in a dry ice ethanol bath for 15 minutes. 3.5ul of proteinase K (10mg/ml) was added to each sample. DNA preps were incubated at 60°C for 30 minutes. 2ul more proteinase K was added and samples were incubated for another 30 minutes at 60°C. To denature proteinase, samples

were incubated at 95°C for 10 min.

PCR was performed in 50ul reactions as described in (Kocher and Wilson, 1991) I tried amplification with each*unc-54* primer (JC32, JC33, JC34, JC35, JC36) with each of the Tc1 primers (JC55 and JC56) (shown in figure 3.1). I also tried PCR with several *unc-54* primers together in a reaction with a single Tc1 primer. The amplification protocol was 30 cycles of 94°C for 30 seconds, 54°C for 1 minute, and 72°C for 2 minutes were used for amplification.

PCR products were electrophoresed on agarose gels and stained with ethidium bromide. Gels were photographed and then transferred by Southern blotting to nitrocellulose membranes essentially as described by Southern, 1975) Radiolabeled probes were prepared by random primed labeling of clones containing the desired target gene. Blots were hybridized (in 50% formamide) with probes overnight at 42°C. Blots were washed twice in 3X, 1X, and 0.3X SSC at 65°C. Blots were exposed on X-ray film and developed several hours later.

Sib-selection PCR with nested PCR to isolate Tc1 insertions in *unc-54* and *unc-22*

Problems with the first method used to isolate new insertions lead to experiments using nested PCR to detect Tc1 insertions. Nested PCR increases the specificity and efficiency of PCR by using a series of two reactions. PCR is performed using a pair of "outer" PCR primers (e.g. JC66 and JC56) and the products from this first reaction are used as templates for a second PCR using a nested set of primers (JC67 and JC58). In theory, it is unlikely that non-specific amplification products from the initial PCR will contain binding sites for the primers used in the nested PCR. Thus, nested PCR adds an additional level of specificity to amplification reactions. In addition nested PCR allows the detection of rare template molecules. Because nested PCR involves two rounds of amplification (as many

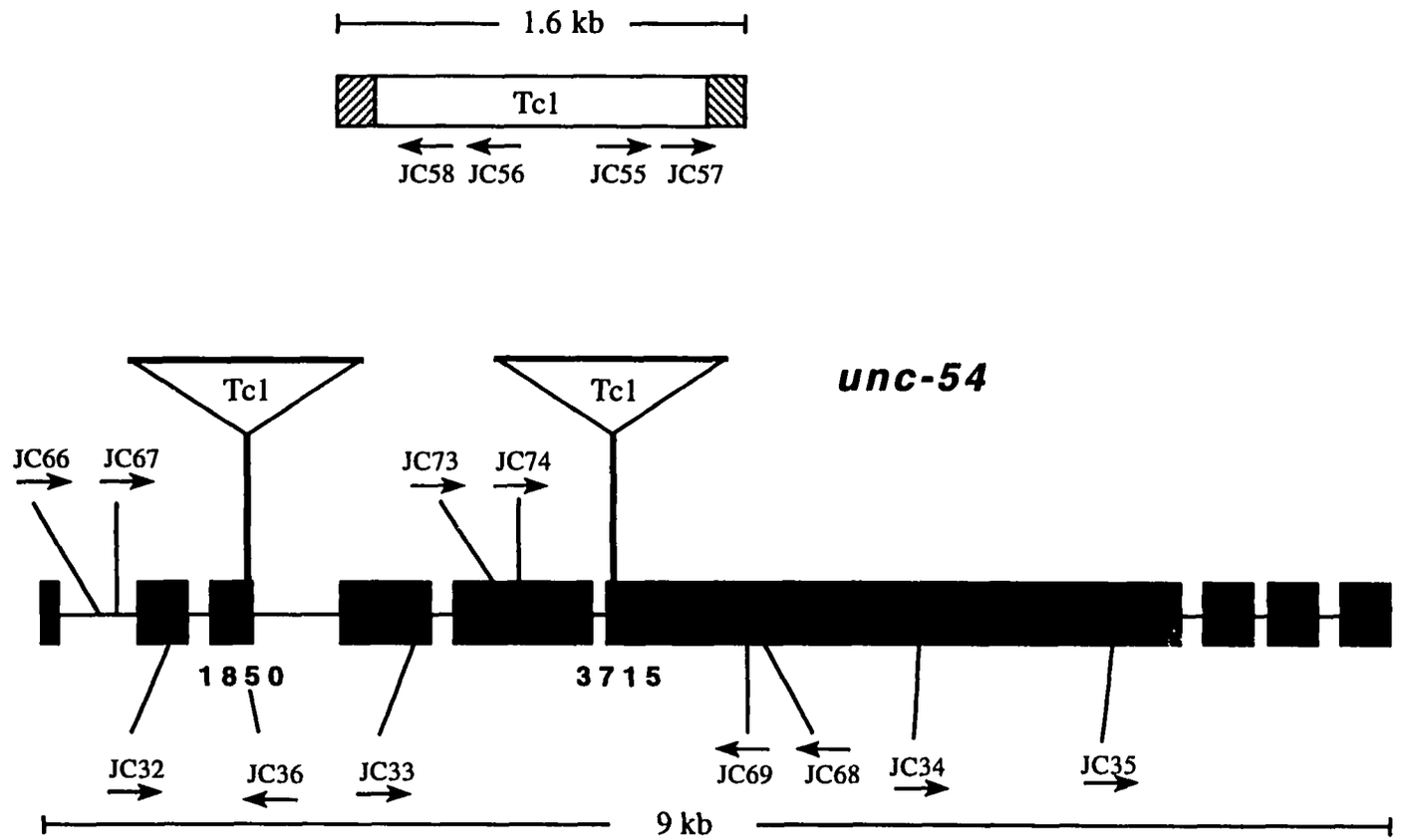


Figure 3.1 Shows the location of *unc-54* and Tc1 primers used in PCR experiments

as 60 thermal cycles), specific and efficient amplification from even single template molecules is possible.

Choice of a strain for sib-selection

TW186, the *mut-2* strain used in the experiments described above grows very slowly. To avoid difficulties in culturing and maintaining the *mut-2* strain for future sib-selection endeavors, I selected a healthier *mut-2* strain, TW332. TW332 was isolated in the same manner as TW186, as a spontaneous *unc-54* revertant of TR674 *unc-54*::Tc1 mut-2 (r459).

Results:

PCR and Southern Blotting to detect insertions

DNA was prepared from 50 populations of TW186 animals each started with approximately 50 worms. PCR amplification was performed using *unc-54* primers JC32, JC33, JC34, JC35 and JC36 with *unc-22* primers JC55 and JC56. In most cases PCR was performed using one *unc-54* primer and one Tc1 primer in a reaction. Regardless of which *unc-54* primer was used or which Tc1 primer was used, all reactions shared one common feature, the presence of multiple products. Generally, DNA from every population of TW332 would produce a similar banding pattern. For some primer combinations amplification resulted in a smear of products when analyzed on an agarose gel. It seemed apparent that simply performing PCR with gene and transposon specific primers was not specific enough to detect new insertion events. To determine which of the numerous amplification products resulted from Tc1 insertion into the *unc-54* gene I transferred the PCR products to nitrocellulose membranes and probed with a cloned copy of the *unc-54* gene (plasmid pUNK-54).

Southern hybridization did little to discriminate between PCR products. Often, most of the PCR products in a lane would hybridize with the probe. Since different populations of

TW186 produced essentially the same set of bands for a particular primer set, all populations shared essentially the same pattern of banding on Southern blots. It seemed unlikely that every band represented a germ-line insertion of Tc1 in *unc-54*. The bands common to every PCR were probably the result of non-specific amplification. Hybridization of the *unc-54* probe to non-specific products might occur because the probe sequence includes the primer sites in *unc-54*. Therefore, any product amplified with an *unc-54* primer might hybridize with the probe and be detected after prolonged exposure. To eliminate the problems associated with nonspecific amplification I chose to focus my attention on the rare PCR products which were unique to particular populations of TW186. Amplification products from three populations were singled out for further analysis.

As described in the methods section, TW186 populations were divided into thirds prior to DNA preparation. One-third of the worms were placed on growth media and maintained for possible sib-selection. DNA was prepared from another one-third of the animals and screened by PCR for *unc-54* insertions. The remaining one-third were kept frozen, pending the results of the first PCR experiment. If the results of the first PCR indicated that a particular population of TW186 contains a new Tc1 insertion, DNA was prepared from the frozen worms corresponding to that population. The DNA is screened with *unc-54* and Tc1 primers to determine if the insertion detected in the first PCR is present in another third of the population. The logic behind such a scheme is as follows. Sib-selection is likely to result in enrichment for insertion containing animals only if the insertions occur in the germline, and only if enough animals containing the insertion are present in the population where insertion is detected. Insertions occurring in somatic tissues cannot be enriched by sib-selection. In addition, insertions occurring late in a culture of TW186 may be present in only one or a few animals and will not be propagated after population subdivision. DNA was prepared from frozen worm samples for the three populations which contain potential *unc-54*::Tc1 insertions. In all three cases PCR

amplification of these second sets of DNA samples,using the primers that detected an insertion in the first set of amplifications, resulted in a failure to amplify the novel band. Since the product did not amplify from the sample of remaining worms, it appeared unlikely that sib-selection would result in enrichment for animals carrying these insertions.

After screening 50 populations with PCR and Southern blotting, a few lessons became clear. First, greater specificity is required; PCR that amplifies numerous nonspecific products from every DNA sample is undesirable. Second, Southern hybridization that does not allow sufficient discrimination between PCR products is clearly unacceptable. Third, and perhaps most importantly, the method chosen to identify new element insertions should be significantly faster than PCR followed by Southern blotting. At the time when DNA is prepared from TW186 populations, plates contain approximately 5000 animals. One-third of these worms are placed on a single petri dish at the time of DNA preparation and allowed to grow. These populations are maintained on plates until PCR and blotting results indicate that a particular population contains a desired insertion. At this point the population is subdivided, cultured and screened for insertions. The problem is that the third of the worm population placed on plates to grow consist of close to 2000 animals. These animals quickly grow to fill the plate. If the culture grows for too long (only a few days) the animals will starve, making the recovery of mutants more difficult.

The next section describes the results of further experiments aimed at isolating new Tc1 insertions. Attempts were made to address and circumvent the difficulties encountered with identification of new insertions with PCR and Southern analysis.

Sib-selection with nested PCR to detect insertion events

Difficulties with the use of PCR and Southern blots to identify new insertions led me to try an alternative protocol to identify insertion events. Zwaal et al. (1993) report the successful application of a nested PCR method to detect new Tc1 insertion events. I used

three nested primer sets in the *unc-54* gene and one nested primer set in the *unc-22* gene to identify new insertions of Tc1 in these genes. Each gene-specific nested primer set was chosen because of its close proximity to sites previously known to be targets for Tc1 insertion. A strain of worms containing a previously isolated germ-line Tc1 insertion was obtained for each primer site in *unc-54* and *unc-22*.

To establish the efficiency and specificity of these nested primer pairs I performed control reactions (referred to as "reconstruction experiments") using templates known to contain transposon insertions. The purpose of these experiments was to determine whether the nested primers could amplify rare insertion-containing templates in a background of non-insertion containing molecules. Three previously characterized germ-line insertions of Tc1 made these experiments possible. I used one *unc-22*::Tc1 allele and *unc-54*::Tc1 alleles r323 and r360, with insertions of Tc1 at positions 1850 and 3715 respectively in the *unc-54* gene.

Worms were collected from 2 populations (that lack an insertion of Tc1 in *unc-54*) containing a total of 10,000 animals each. To one of these populations a single TR656 animal was added to the population of insertion lacking animals. To a second population, ten TR656 animals were added. DNA was prepared from both populations and PCR was performed using outside primers JC73 and JC55. Products from the first PCR were diluted and used in nested amplification reactions with primers JC74 and JC75. Nested PCR results in specific amplification from the insertion containing template even when it is present among a 10^4 excess of wild-type templates. By these criteria, all primer sets appeared to be adequate for detection of rare insertion containing templates from populations of *C. elegans*.

unc-22 is a difficult target for detecting new insertions by PCR

DNA was prepared from 30 populations of TW332 each started with approximately 50

animals. Cultures were grown until there were approximately 5000 animals on the plate. DNA was screened with nested PCR primers in *unc-22* and Tc1. Amplification reactions from many populations of TW332 contained products representing potential insertions of Tc1 into *unc-22*. Assuming that these insertions are germ-line insertions of Tc1 in *unc-22*, the next step in the sib-selection/PCR protocol would be sub-division of worms from populations generating an PCR product. Given the large number of potential insertions detected by PCR, this step would have meant committing to hundreds of DNA preparations and thousands of amplification reactions. To ensure that the PCR products represented insertions into the expected target region of *unc-22*, I sequenced several independent PCR products. The sequences revealed a problematic and unexpected result of the PCR experiment. None of the products corresponded to insertion into the expected region. Upon closer inspection I realized that the *unc-22* primers chosen for the sib-selection experiment contained multiple mispriming sites within the *unc-22* gene. *unc-22* is an extremely large gene by *C. elegans* standards, spanning more than 60kb on linkage group IV. The *unc-22* gene and gene product contain highly repetitive structural features. The primers used for PCR are contained within one of these repetitive motifs and result in amplification from a number of positions in the *unc-22* gene. Although germ-line Tc1 insertion occurs in this gene, it is difficult to target insertions to a particular gene region due to its highly repetitive structure.

unc-54::Tc1 insertions are detected by PCR but are difficult to isolate by sib-selection

To alleviate the problems caused by mispriming within a gene we designed two nested sets of primers for regions of the *unc-54* gene, careful to avoid repetitive sequences in the *unc-54* gene. In total, approximately 300 primary cultures of TW332 were screened by PCR using these *unc-54* primers. Initial rounds of screening were performed on 20 populations at a time. Although it is feasible to screen a larger number of populations at a

time, potential insertions detected in the initial round of screening require numerous rounds of enrichment by sib-selection to isolate animals homozygous for an insertion containing allele. Screening a large number of populations in an initial round of screening would lead to an unmanageable number of populations in subsequent rounds of sib-selection.

One Tc1 insertion containing allele was successfully obtained using the sib-selection PCR protocol. A flow chart is shown (figure 3.2) describing the process of screening and population subdivision leading to isolation of the insertion. In addition to this successful isolation of an insertion-containing strain, there were numerous insertions which were detected in early rounds of sib-selection but were lost in successive rounds. Insertions detected by PCR but not enriched by sib-selection probably arise for two reasons. First, insertions occurring late in the culture will be present in only one or a few animals and are likely to be lost during sib-selection. Second, insertions occurring in somatic tissues of animals will be detected by PCR but not inherited and hence not enriched by sib-selection.

A population of animals was isolated by sib-selection in which DNA from every single individual would amplify a 1100bp product with primers JC69 and JC58. This product was sequenced with the expectation that it would represent a Tc1 insertion in *unc-54*. Surprisingly, the sequence of the PCR product, although very similar to *unc-54*, is better interpreted as an insertion of Tc1 into another *C. elegans* gene, *myo-1*! The strain containing this insertion, TW386, has a Tc1 element inserted at position 5938 in *myo-1*, in an intron. This Tc1 insertion results in no obvious phenotype.

How did I isolate a *myo-1*::Tc1 insertion using *unc-54* primers? The answer lies in analysis of myosin genes in *C. elegans*. *C. elegans* contains 4 genes encoding sarcomeric myosin heavy chains (MHCs); *myo-1*, *myo-2*, *myo-3*, and *unc-54* (Dibb et al., 1989). The nucleotide sequences of these genes share many similarities, as do their protein products. A particularly well conserved region of these genes is located in the region containing PCR primers JC68 and JC69. Twenty-one bases out of 22 in *myo-1* are

Round 1

⇓

20 initial populations each containing approximately 5000 animals were screened with primers JC68 and JC69. One out of twenty of these populations generated a PCR product of ~1100bp.

⇓

SUBDIVISION

⇓

Round 2

⇓

10 populations were started with approximately 500 worms per plate (all of the worms remaining from the population where an insertion was detected). Worms were grown and tested for the presence of the insertion. 5 out of 10 populations were positive.

⇓

SUBDIVISION

⇓

Round 3

⇓

30 populations of 50 animals each were seeded with worms from 1 of the 10 populations from the last round. Worms were grown and tested for the presence of the insertion. 15 out of 30 populations were positive.

⇓

SUBDIVISION

⇓

Round 4

⇓

10 populations of 10 worms each were started from 1 of the 15 positive populations from the last round. An additional 30 populations were started with single animals from the same population. Worms were grown and tested for presence of the insertion. All ten populations containing started with 10 animals were positive for the insertion. 14 out of 30 of the populations seeded with a single individual were positive for the insertion.

⇓

Round 5

⇓

DNA was prepared for 20 single worms from one of the positive populations started with a single animal. All twenty animals screened contained the insertion suggesting that this strain is homozygous for a germ-line Tc1 insertion.

Figure 3.2 Flow chart for detection and sib-selection of an insertion detected with JC68 and JC69.

identical to primer JC68 and 19 out of 22 bases are identical to JC69. All differences between the *myo-1* sequence and the primers designed for *unc-54* lie at least 8bp away from the 3' ends of the primers. Thus primers JC68 and JC69 are likely to detect insertions in *myo-1* as well as *unc-54* (and likely the rest of the MHC gene family). This problem is exacerbated by the use of nested PCR. Nested PCR is useful for eliminating non-specific amplification products produced in the first round of PCR by requiring that products for nested PCR amplification contain priming sites for nested primers. The isolation of a *myo-1*::Tc1 insertion points to an unexpected complication imposed by nested PCR. Multigene families by definition, contain conserved sequences. If PCR primers are designed for a conserved region of the gene family, they may prime amplification of products from different genes. Differences between the gene sequence and the PCR primer used in the initial reaction are expected to reduce the efficiency of amplification. However, if the template generated in the first PCR also contains primer binding site for the nested primer (as might be expected from a multigene family) the product may be amplified exponentially in a nested PCR.

A new set of nested *unc-54* primers were carefully designed, avoiding not only repetitive regions within the *unc-54* gene, but also regions where the nucleotide sequence is conserved between members of the myosin heavy chain gene family. Attempts at sib-selection/PCR with these primers revealed yet another complication associated with the technique. 20 populations were screened with nested *unc-54* primers JC66 and JC67. Surprisingly, a ~440bp product was amplified from every population of TW332 screened. Characterization of this common PCR product is described extensively in CHAPTER IV and will not be discussed here. For the purposes of the sib-selection experiments, this common PCR product was ignored. Only insertion products greater than or less than 440bp were selected for enrichment by sib-selection. Many PCR products of this sort were detected, but none were successfully enriched by sib-selection. In several cases a particular

insertion was detected in several rounds of sib-selection but ultimately lost in later rounds of population subdivision.

Two explanations for detecting insertions in populations without successful enrichment by sib-selection were discussed earlier. Either insertions arise late in the culture and are not present when populations are subdivided, or insertions occur in somatic tissue and are not inherited. The large number of potential insertion events which were detected in several rounds of sib-selection but ultimately not enriched are likely due to somatic insertion events into sites in *unc-54*. Germ-line insertions may be lost if they occur late in the culture. If they are lost, it is unlikely that a PCR product consistent with such an insertion would be detected after several rounds of sib-selection. Somatic insertion events on the other hand might be detected in several rounds of sib-selection. During the later rounds of sib-selection, populations are subdivided and new cultures are started using a smaller number of worms than in the previous round of subdivision. A result of this procedure is that a smaller number of progeny are present on the plates before DNA is prepared and screened by PCR after each round of sib-selection. In early rounds of sib-selection there are approximately 5000 animals on a plate when DNA is prepared. In later rounds of sib-selection there may be only hundreds or even tens of worms on a plate when DNA is prepared. If a particular insertion into *unc-54* arises in somatic cells at a frequency of ~ 5×10^{-5} we might expect to detect an insertion about once in every ten populations screened. A somatic insertion of this type is expected to be detected in early rounds of sib-selection when DNA is prepared from a large number of animals, but rarely detected in DNA prepared from a small number of animals.

At this point I decided that it was unlikely that the sib-selection PCR protocol as outlined above would provide the means necessary to isolate a large number of germ-line Tc1

insertions in *unc-22* or *unc-54*. The surprisingly high frequency of somatic insertion into *unc-54* lead me to investigate this aspect of Tc1 activity in greater detail, as described in the next chapter.

CHAPTER IV

HIGH FREQUENCY SOMATIC INSERTION OF TC1 IN *C. ELEGANS*

Summary:

Transposition is a regulated process. For some transposons this regulation responds to developmental stage or cell type. In *C. elegans*, previous work has shown that excision of the transposon Tc1 is 1000-fold more frequent in somatic cells than in the germline. I have discovered that insertion of Tc1 also occurs at remarkably high frequency in the soma. In the most dramatic example, insertion of Tc1 was detected at the same site in the *unc-54* gene in nearly every animal screened. This site was previously shown to be a "hotspot" for germ-line insertion, although at a frequency several orders of magnitude less than the levels now detected. I believe these insertions are somatic events because they increase in frequency during development but are not transmitted to progeny based on both genetic and molecular evidence and because I detect them in animals lacking a germline. Additional sites in *unc-54* and *src-1*, another *C. elegans* gene, were identified as frequent targets for insertion of Tc1; however, none are hit as frequently as the *unc-54* "hotspot". Somatic insertion of Tc1 depends on genetic background; it occurs at very high frequency in several wild-type genetic backgrounds and the *mut-2* mutant background of *C. elegans*, but not in the wild-type strain Bristol N2. These results are important for understanding the evolution of mechanisms involved in regulation of transposon activity, and for the use of Tc1 as a tool for reverse genetic approaches in *C. elegans*.

Introduction:

Eukaryotic genomes are replete with transposable elements. Insertion and excision of transposable elements can generate changes in gene sequence, gene expression, and chromosome structure (reviewed in Berg and Howe, 1989; Lambert et al., 1989). Understanding the role transposon-generated genetic variation has played in genome and organismal evolution requires characterization of the rates, patterns, mechanisms and phenotypic consequences of transposable element activity. If transposition occurs frequently, and the majority of insertion and excision events are severely deleterious, individuals harboring these elements may suffer a selective disadvantage and be eliminated from the population. This process would lead to the eventual loss of the transposon from the population. In light of the potential consequences of unchecked transposition, it is not surprising that mechanisms exist to regulate when, where and how transposons move and to mitigate the effects of their insertion.

Transposons are often considered a type of selfish DNA. They persist because they make additional copies of themselves in the genome, not because of any specific contribution to the phenotype of their hosts. If we assume that there is competition among element families for sites in the genome, elements that replicate efficiently in the germline will eventually replace elements that do not (Orgel and Crick, 1980). Replication in the soma, on the other hand, is not expected to increase the probability of long-term persistence of a transposon. In fact, somatic activity of an element might have deleterious effects on cells containing them. If the deleterious consequences of insertion in somatic cells affects the "host" organism, it may lead to a decrease in the probability of long-term persistence of a transposon. Some transposons do not transpose in somatic tissues. For example, P element transposition in *Drosophila* is restricted to the germline due to tissue-specific splicing of the P element-encoded transcript (Laski et al., 1986). However, somatic transposon activity is observed for many different elements, often at levels far exceeding

those of the germline. For example, Tc1 elements in *C. elegans* undergo low levels of excision in the germline, (Eide and Anderson, 1985; Moerman et al., 1986) but excise at much higher frequency in somatic cells (Emmons and Yesner, 1984; Eide and Anderson, 1988). Similar observations have been made for mariner elements in *Drosophila* (Bryan et al., 1987) and *Mu* elements in maize (Doseff et al., 1991). Somatic activity may arise simply because the factors necessary for germ-line transposition of some elements are not confined to the germ cell lineage. If somatic transposition is selectively neutral, replication of elements in somatic cells might arise as a simple property of selfish DNA (i.e., they replicate in somatic cells because they can). Elements that replicate more efficiently in somatic cells will be found at higher copy number in somatic cells than elements that cannot. Understanding how transposons are differentially regulated in the germ and soma may help clarify these issues.

Tc1 is active in both germ-line and somatic tissues (Eide and Anderson, 1985) however, regulation of Tc1 activity differs in these two tissue types (Emmons et al., 1986). Collins et al. (1987) isolated "mutator" mutants that exhibit elevated levels of germ-line excision without affecting frequencies of somatic excision, suggesting that Tc1 regulation is tissue specific. Mutator mutants also exhibit a significant increase in the frequency of germ-line transposition events suggesting that insertion and excision (at least in the germline) are regulated by common factors. Germ-line activities of *C. elegans* transposons Tc3, Tc4 and Tc5 are also elevated in the mut-2 background (Collins et al., 1989; Yuan et al., 1991; Collins and Anderson, 1994).

Germ-line transposition and excision of Tc1 is detectable in the Bergerac strain of *C. elegans* (Eide and Anderson, 1985a) but not in the Bristol (N2) strain (Eide and Anderson, 1985b). The Bergerac genome harbors approximately 500 copies of Tc1 compared to 26 copies in the Bristol genome (Emmons et al., 1983; Rosenzweig et al., 1983; Egilmez et al., 1995). Unlike the difference in frequency of germ-line excision, levels of somatic

excision of Tc1 are comparable between these strains (Harris and Rose, 1986; Eide and Anderson, 1988). Thus, in the Bristol genome, elements are competent to move but appear to be suppressed in germ-line tissue. The availability of strains with different Tc1 copy numbers and varying levels of element activity have proven useful for transposon tagging efforts in C. elegans (Moerman et al., 1986).

As discussed in CHAPTER I, the ability of transposons to insert at new sites has lead to their exploitation as tools for molecular geneticists. One complication in isolating animals containing germ-line transposon insertion and excision products is somatic transposon activity (as discussed in the previous chapter). When identifying new insertion or excision products using PCR (the method of choice in *C. elegans*, Rushforth et al., 1993; Zwaal et al., 1993), the products of somatic insertion and excision may be indistinguishable from their germ-line counterparts. This can lead to a serious problem of false positives when screening for new germ-line insertion and excision events. Knowing the relative rates of transposon activity in the germline and soma allows the design of more efficient screens for desired products of transposon movement. This information is also important for understanding the evolution of transposable elements.

To understand the evolutionary history of transposons and predict their mutagenic potential it is necessary to know the spectrum of different mutations induced by transposon insertion and excision and the rates at which they occur. As discussed in the previous chapter, one difficulty in interpreting frequencies of transposon insertion and excision is the tendency for most genetic methods (that rely on phenotype to detect transposon movement) to underestimate the true level of activity.

Some methods used to detect transposon movement, such as *in situ* hybridization of element probes to *Drosophila* polytene chromosomes, can be used to identify new insertions without regard for the mutant phenotype and can allow detection of insertion over a broad range of sites. Data regarding the distribution of transposon sequences in the

Drosophila genome (Charlesworth et al. 1992) as well as estimates of the rates of germ-line insertion and excision (Nuzhdin and Mackay 1995) have been determined for a variety of element families. *In situ* hybridization does not, however, allow fine scale analysis of insertion sites at the DNA sequence level. Hence, little information is available to compare the differences in frequency of insertion into distinct portions of the genome e.g. gene vs. intergenic, intron vs. exon, promoter vs. coding region.

Transposon insertion generates a distinct molecular structure, namely the insertion of transposon DNA into a target. This molecular structure can be used to identify new insertions without regard for the phenotype associated with the insertion. We used a PCR based approach similar to one described previously for *Drosophila* (Ballinger and Benzer, 1989; Kaiser and Goodwin, 1990) and *C. elegans* (Rushforth et al., 1993; Zwaal et al., 1993) to detect new Tc1 insertions into the *C. elegans unc-54* gene and have identified sites which are frequent targets for somatic insertion of Tc1. I know that these insertions are somatic since they can be detected in animals lacking germ tissue. One site is hit so frequently that almost every animal contains an insertion of Tc1 at precisely the same nucleotide position. This hotspot for somatic insertion resides at the precise location of a hotspot for germ-line transposition.

Materials and Methods:

C elegans strains and maintenance:

Worms were cultivated on agar plates seeded with *Escherichia coli* strain OP50 (Brenner, 1974). Strain TW332 *mut-2*(r459) was isolated in our laboratory as a spontaneous wild-type revertant of TR674 *mut-2(r459)*; *unc-54*(r323). TR674 as well as TR1299 *unc-54* (r323) were obtained from Phil Anderson. Wild isolates of *C. elegans* EM1002, N2, TR403 and DH424 were obtained from the *Caenorhabditis* stock center. All of these strains were grown at 20°C. A strain carrying the temperature sensitive *glp-*

4(bn2) allele (Beanan and Strome 1992) was provided by Susan Strome. The permissive temperature for this strain is 16°C and the restrictive temperature is 25°C. Genetic manipulation of strains was performed as detailed by Brenner (1974).

DNA extraction and PCR amplification:

DNA from single animals was extracted by placing one worm in a microfuge tube containing 30ul WLB and 1ul proteinase K (10mg/ml). DNA from groups of 10 worms were prepared by placing 10 animals in 50ul of lysis buffer with 1ul proteinase K. Extractions were frozen for 15 minutes in dry ice/ethanol bath and then incubated at 65°C for 1 hour and then heated to 95°C for 10 minutes.

Nested PCR amplifications were performed using several primer sets. The names and sequences of primers are shown in Table 4.1 and their locations are shown in Figure 4.1.

Table 4.1 Sequences of PCR primers used to detect Tc1 insertions.

primer name	specific for gene:	sequence 5'-->3'
JC56	Tc1	GCTGATCGACTCGATGCCACGTCG
JC58	Tc1	TTGTGAACACTGTGGTGAAGTT
JC66	*unc-54*	TTAGACCATTTTTCAACACAAG
JC67	*unc-54*	CTGAATTCTGATCTCTTTTGTA
JC73	*unc-54*	AAATCTACTCTGACTTCCGT
JC74	*unc-54*	TTGCCAATCAAGGATTACTG
JC60	*src-1*	GTCAACTTACATTCCCAGCACCTC
JC61	*src-1*	TCGTGCCTCGTAAATGTCCTCTTC

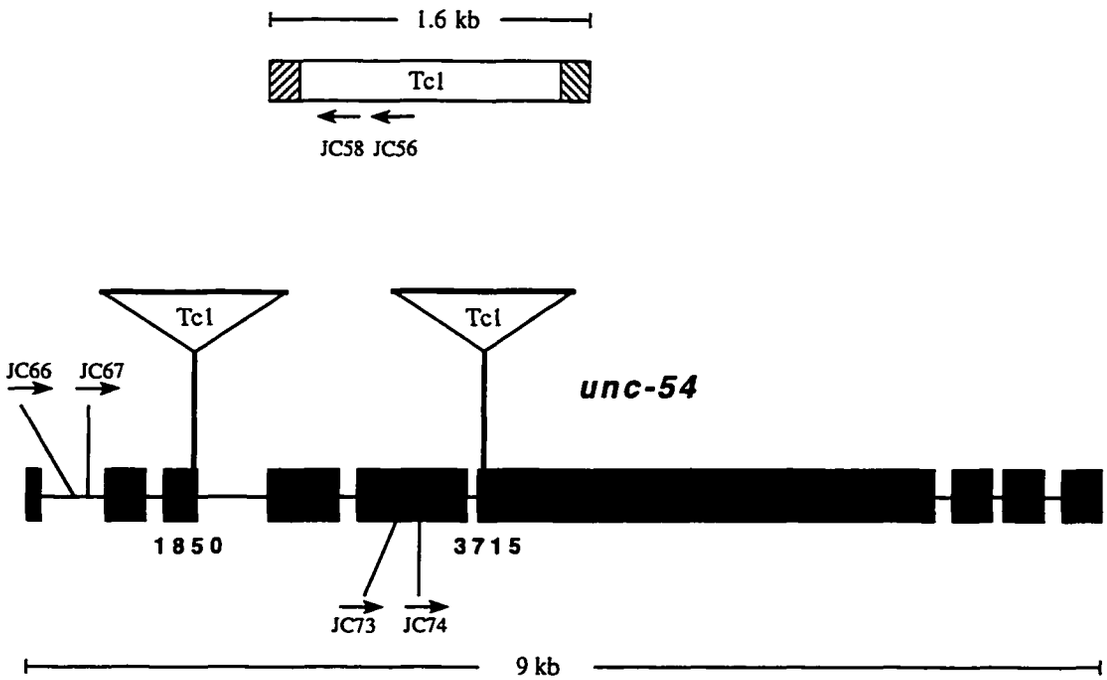


Figure 4.1 Location of PCR primers in *unc-54* gene and Tc1 transposon. PCR amplification with gene and transposon specific nested primer sets occurs only when Tc1 inserts in close proximity to the *unc-54* primer sites. The position of sites 1850 and 3715, identified as hotspots for insertion of Tc1 into *unc-54*, are shown in the illustration.

For most experiments, 5ul of template DNA from single worms (approximately one-sixth of a worm's DNA) or groups of 10 worms was added to each reaction. The entire 30ul of templates prepared from single ablated TW332 animals and their unablated controls was used in PCR. Amplification reactions were performed essentially as described in Kocher and Wilson (1991). 50ul reactions were subjected to 30 cycles of 94°C for 30 seconds, 54°C for 1 minute and 72°C for 2 minutes. PCR products from the initial reaction were diluted 1:10 with H_2O and 1ul was used as template in nested amplification reactions with the same conditions described above. PCR products from the nested amplifications were visualized on 1% Seakem agarose gels stained with ethidium bromide.

Genomic Southern blots:

DNA was prepared from strains TW332, N2 and TR1299 as previously described (Eide and Anderson, 1985). DNA was cut with *BamH1* and electrophoresed through 1% agarose gels. DNA samples were blotted to nitrocellulose membranes essentially as described by Southern (1975). Membranes were hybridized overnight with a ^{32}P radiolabeled *unc-54* plasmid, punk-54. Plasmid DNA was labeled by primer extension of random hexamers as described by the manufacturer (Amersham).

Detection of insertions in parents and their offspring:

Single adult hermaphrodites were allowed to lay eggs and then DNA was extracted from the "parental" worm. Several L1 larval progeny hatching from these eggs were collected and two days later adult progeny were picked from the plate. DNA was prepared from the larval and adult progeny and used as template in nested PCR with primers JC66 and JC67 in *unc-54* and JC56 and JC58 in Tc1.

Sequencing of PCR products:

PCR products were sequenced directly by cutting bands of interest from Nusieve low-melt agarose (FMC) gels. Gel slices were melted by incubation at 65°C for 10 minutes. 10 units of agarase (SIGMA) was added to melted gel bands and then incubated at 37°C for one hour or until agarose was digested. Digested gel bands were used as templates in cycle sequencing reactions containing dye-labeled dideoxy terminators (ABI). Extension products were purified through a Sephadex column. Purified sequencing products were run on an ABI 373A automated DNA sequencer.

Construction of strains that contain somatic Tc1 activity and the *glp-4(bn2)* allele:

glp-4(bn2) animals are temperature sensitive sterile mutants. When worms are raised at the restrictive temperature (25°C), germ nuclei fail to proliferate resulting in adult animals severely depleted of germ nuclei. TW332 and Bergerac hermaphrodites were mated with males heterozygous for *glp-4(bn2)*. F1 animals were plated singly and allowed to lay eggs. Several F2 animals from each F1 plate were picked and plated singly. Approximately twelve F3 L1 larvae were picked from each F2 plate, placed on plates and shifted to growth at 25°C. F2 clones that gave rise to F3 progeny which were sterile at 25°C (and hence *glp-4(bn2)* homozygotes) were retained. For each strain which was potentially homozygous for *glp-4*, several single worms were raised at 16°C, picked and screened for insertions of Tc1 in *unc-54* using nested PCR primer pairs JC56 and JC58 and pairs JC66 and JC67. Strains which contained worms producing a PCR product were retained. These new strains contain both the temperature sensitive *glp-4(bn2)* allele and a high level of Tc1 activity at 16°C. Worms from this new strain were raised at 25°C and DNA was prepared from single animals and subjected to nested amplification with the *unc-54* and Tc1 primers.

Laser ablation of TW332 larvae:

Early TW332 L1 larvae were picked onto agarose pads and immobilized in a 50mM solution of sodium azide. Worms were visualized under Nomarski interference optics. Z2 and Z3 germ-line progenitor cells were identified and ablated using a laser microbeam. Worms were removed and cultured for 5 days. Worms were picked into lysis buffer, DNA was prepared and then amplified by PCR using *unc-54* and Tc1 primers as descibed above.

Results:

Tc1 insertion into the *unc-54* gene occurs frequently in TW332

We used a modification of the procedures described by Rushforth et al. (1993) and Zwaal et al. (1993) to detect new transposon insertions in DNA prepared from populations of *C. elegans.* This technique relies on the fact that a nested set of gene-specific and transposon-specific primer pairs will specifically amplify the junction between gene and transposon sequences. Primers JC66 and JC67 are specific for a region of *unc-54* and were used with Tc1 primers JC56 and JC58 in the PCR (Figure 4.1). To increase the likelihood of observing insertion events we used a strain of *C. elegans*, TW332, that harbors the *mut-2* mutator. This factor is known to increase levels of germ-line insertion and excision of Tc1 (Collins et al, 1989). Unexpectedly, every population of TW332 (each containing approximately 5000 animals) screened by PCR contained an insertion of Tc1 at the same or nearly the same site (based on the migration of products on an agarose gel). We screened smaller and smaller populations of TW332 and eventually single animals to investigate this phenomenon .

Amplification of junctions between *unc-54* and Tc1 from single TW332 animals reveals that approximately 70% of adult worms contain an insertion at or near the site

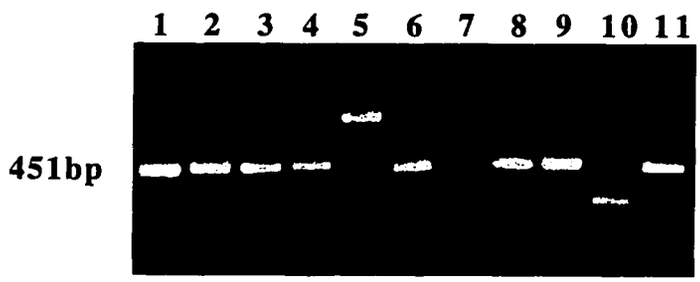


Figure 4.2 This agarose gel shows typical PCR products amplified from single animals using the nested *unc-54* primer JC67 and the nested Tc1 primer JC58. Lane 1 contains a 451 bp product amplified from a single TR1299 animal known to contain a Tc1 insert at position 1850 in *unc-54*. Lanes 2-11 are each products of amplification from a single TW332 adult hermaphrodite.

shown to be a "hotspot" for germ-line insertion of Tc1 (Eide and Anderson, 1988). Sequences of several independent PCR products, discussed below, reveal that many of these insertions are at the hotspot (position 1850, numbered as in Karn et al., 1983), a TA dinucleotide 3 bp upstream of the 5' splice site of the *unc-54* third intron (see figure 4.1). Figure 4.2 (lanes 2-11) shows an agarose gel of typical PCR products from single TW332 animals amplified with primers JC67 and JC58 (see figure 4.1). Figure 4.2 lane 1 shows the JC67 and JC58 PCR product amplified from a single animal of strain TR1299 [genotype*unc-54(r323)*] using the same primers. *r323* contains a germline insertion of Tc1 at the hotspot. In Figure 4.2, eight or nine out of 10 single TW332 animals produce a band of the same size (451 bp) as *r323*. Other bands were observed in some reactions (e.g. lanes 5 and 10). Table 4.2 summarizes the frequency of insertion into the hotspot for 45 single worms. Thirty-three out of 45 have an insertion at the hotspot. In addition, 18 out of 45 have bands consistent with insertions at other sites in the region and 10 of these 18 also have an insertion at the hotspot. Of the 18 other products amplified from single animals, 10 are detected in animals that also generate the 451bp product (e.g. lane 10). Collectively, these results indicate a very high frequency of Tc1 transposition, especially considering that I have examined only one part of one gene.

Eide and Anderson (1988) isolated 11 spontaneous Tc1 induced *unc-54* germ-line mutations in *C. elegans* strain Bergerac. 7 out of 11 insertions occurred at the hotspot. Animals homozygous for an insertion of Tc1 at this position exhibit a typical *unc-54* loss-of-function phenotype; worms are paralyzed, flaccid, and egg-laying defective (Eide and Anderson, 1988). None of the TW332 animals containing insertions at the hotspot detected by PCR had the mutant phenotype expected for a germ-line insertion of Tc1 into *unc-54* coding sequence. The lack of a mutant phenotype and the high frequency of insertion suggest that the insertions we detect might be occurring in somatic cells and apparently do not affect a large number of muscle cells.

Table 4.2 Summary of somatic insertion frequencies in different strains and life stages. The strain names are followed by the life stage of the animal(s) considered. All animals were adults, unless otherwise indicated. A strain name followed by "pop10" refers to a DNA sample prepared from ten animals.

strain	# animals or populations screened	# insertions into hotspot	frequency
TW332 adults	45	33	0.73
TW332 L1 larvae	40	1	0.03
EM1002 adults	55	35	0.64
EM1002 L1 larvae	50	1	0.02
DH424	40	4	0.10
TR403	40	4	0.10
MT3126	25	14	0.56
N2	25	0	-
TW332 pop10	10	10	1.00
TW332 L1 pop10	20	5	0.25
EM1002 pop10	10	10	1.00
DH424 pop10	10	10	1.00
TR403 pop10	10	9	0.90
N2 pop10	10	1	0.10

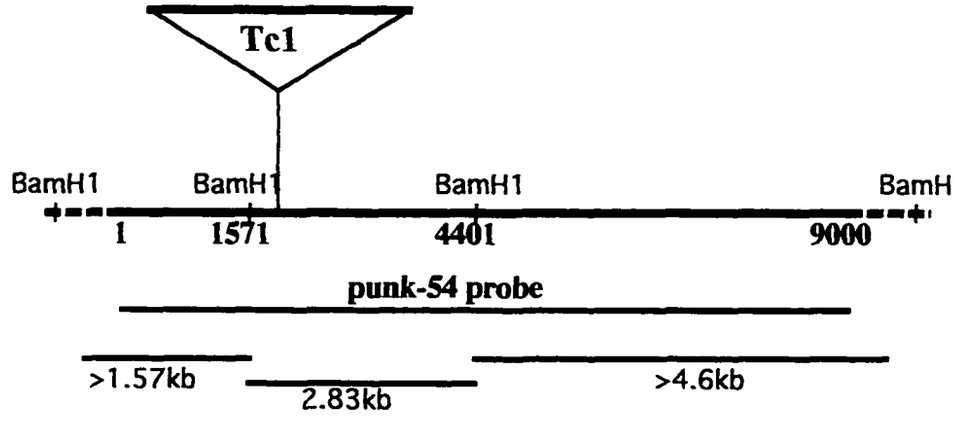
The Tc1 primers used to estimate the frequency of insertion into *unc-54* anneal within the unique portions of Tc1 (i.e. not within the inverted repeats). Therefore, these primers detect Tc1 insertions occurring in only one orientation. Tc1 is known to insert in both orientations and sites that are frequent targets for insertion of Tc1 in one orientation are also targets for insertion in the opposite orientation (van Luenen and Plasterk, 1994). I amplified *unc-54*::Tc1 insertional junctions from the same 45 single TW332 animals described above using hotspot primers JC66 and JC67 and a set of primers that are specific for the other side of Tc1. I observed insertion into the hotspot at comparable frequencies for insertion in this orientation (data not shown). This suggests that insertion is equally likely in either orientation and that many TW332 animals contain more than one insertion into *unc-54*. Detection of Tc1 insertions in both orientations from a single animal is likely only if the insertions are present in different copies of *unc-54* . It is concievable that two insertions could occur in the same copy of *unc-54,* but only the insertion proximal to the *unc-54* primers would be detected after PCR. The frequencies of insertion into the hotspot reported in Table 4.2 are probably underestimates. Insertion is likely to be at least twice as frequent since insertions into the hotspot are detected in both orientations at approximately equal levels.

A potential explanation for detecting a Tc1 insertion at the same position in almost every animal is that TW332 contains a germ-line insertion of Tc1 at this site. This is unexpected since TW332 was isolated as a spontaneous *unc-54+* revertant of TR674 (*unc-54* (*r323*::Tc1)). TR674 animals are paralyzed because they contain a germ-line insertion of Tc1 at the hotspot. The phenotypic change associated with TW332 (reversion) was assumed to result from Tc1 excision from *unc-54.* However, it is also possible that reversion occurred without loss of the element. This has been observed for other transposons including Tc3 in C. elegans. We observed phenotypic reversion of an *unc-22*::Tc3 mutant without element loss (Mills, 1993). In these cases, slight alterations in the

sequence of the insertion-containing allele altered the consequences of splicing of Tc3 from gene transcripts, leading to the production of an in-frame, functional mRNA. Tc1 is also known to be spliced from transcripts of genes into which it has inserted, (Rushforth and Anderson, 1996) so it is possible that reversion of TR674 is due to a change in the sequence of the *unc-54*::Tc1 allele that alters RNA processing and leads to the production of a functional gene product without loss of Tc1.

To determine if Tc1 is present at the hotspot in *unc-54* in this strain we performed a total genomic Southern blot probed with radiolabeled punk-54, a clone containing the *unc-54* region. The blot is shown in figure 4.3. DNA was prepared from TW332, the wild-type strain Bristol (N2) and TR1299 (a strain containing a germ-line insertion of Tc1 at the *unc-54* hotspot) and digested with *BamH1*. Lane 1 contains DNA from Bristol and a 2.8 kb restriction fragment contains the *unc-54* hotspot region. Lane 2 is TR1299 DNA and contains a faint 2.8 kb fragment and an additional band of 4.4 kb representing the Tc1 insertion at the hotspot. Lane 3 contains TW332 DNA and clearly indicates a 2.8 kb band demonstrating that this strain does not contain a germline insertion of Tc1 at the hotspot.

Tc1 excision products are known to account for approximately 1-5% of filled sites in strain TR1299 making them detectable on Southern blots (Eide and Anderson, 1988) as demonstrated by the faint 2.8 kb fragment seen in TR1299 DNA (Figure 4.3 lane 2) . The ability to detect Tc1 in *unc-54* from almost every single TW332 worm by PCR combined with the fact that a Tc1 insertion is undetectable on Southern blots suggests that the insertions are occurring in somatic tissue. It further suggests that less than 1% of the copies of *unc-54* contain the insertion since a higher percentage of insertion-containing molecules would be detectable on the Southern blot. These insertions probably occur during post-embryonic development since somatic mutations occurring early in development could be propagated in somatic cell lineages and rise to levels greater than 1%.



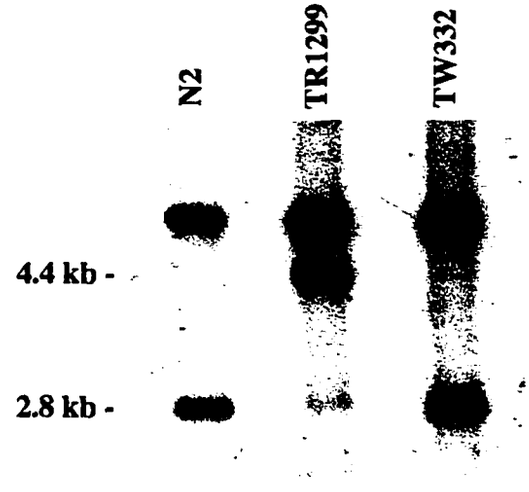


Figure 4.3 Genomic Southern Blot of BamH1 digested DNAs from N2, TR1299 and TW332 worms and probed with punk-54, a cloned copy of unc-54. A BamH1 restriction map is shown above the blot. BamH1 cuts twice in unc-54 at positions 1571 and 4401 generating a 2830 bp fragment and twice in regions flanking unc-54. BamH1 does not cut in Tc1. The punk-54 probe covers the entire length of unc-54, but no flanking sequences, and detects three fragments in wild-type worms. TR1299 is known to contain a 1610 bp Tc1 insertion at position 1850 in unc-54 resulting in a 4440bp BamH1 fragment. A 2830 bp fragment in TR1299 arises from somatic excision of Tc1 from unc-54. Figure 3. PCR products from 10 single TW332 adult animals amplified with nested primers JC58 and JC67.

The PCR results described above indicate that nearly every animal contains at least one insert in their soma making them genetic mosaics for wild-type *unc-54* and *unc-54*::Tc1.

The high frequency of Tc1 insertion into *unc-54* occurs in most wild-type genetic backgrounds

The evidence above shows that a high frequency of Tc1 insertion into the unc-54 hotspot occurs in a *mut-2* mutant background. *mut-2* is known to increase the frequency of germline Tc1 insertion and excision but does not to affect somatic activity (measured as excision, Collins et al., 1987). I wanted to know if high frequency insertion of Tc1 into the hotspot is unique to the *mut-2* mutant background.

Single adult worms and pools of ten worms, from a variety of strains, were screened by PCR to detect insertions of Tc1 into *unc-54*. TW332, Bergerac, DH424, and TR403 all show high levels of insertion into the hotspot in *unc-54* (Table 4.2). In addition, each of these strains contain individuals with insertions at other sites in this region of *unc-54*. In contrast, insertion into this region of *unc-54* is undetectable in single Bristol worms. When pools of ten worms were screened, only one out of ten populations contained an insertion at the hotspot whereas almost every population of the other strains contained an insertion. Frequencies closer to 1 hotspot insertion per PCR were observed only when templates consisted of DNA from several thousand N2 worms (data not shown). Insertions detected at a level of one in several hundred or several thousand Bristol animals is still orders of magnitude greater than the frequency of germ-line insertion into this site. We assume that the insertions detected in Bristol as well as the frequent insertions seen in TW332, Bergerac, DH424, and TR403 occur in somatic cells.

Tc1 insertions arise during culture of TW332 and EM1002, and are not inherited

If the Tc1 inserts I detected indeed occur in somatic cells they should accumulate during

development but not be inherited. To test these predictions I performed an experiment that monitored the presence of a Tc1 insertion in *unc-54* in TW332 parents and their progeny. Single adult hermaphrodites were placed on plates, allowed to lay eggs for 36-48 hours, then picked singly and placed in lysis buffer for DNA preparation. Embryos were allowed to hatch and harvested for DNA preparation in two groups, several L1 larvae were picked singly into lysis buffer, the remaining larvae were allowed to complete post-embyonic development and were collected as adults. All DNA samples were screened by PCR for the presence of Tc1 at the *unc-54* "hotspot" region. Insertions were detected in most "parent" worms (Fig. 4.4 lane 2), very few were detected in larval offspring (Fig. 4.4 lanes 3-7), and most adult offspring contain the insertion (Fig. 4.4 lanes 8-12). In some cases adult progeny contain bands that were not observed in the parent (e.g. lanes 9 and 11). Additionally, 2 out of 5 TW332 parents lacked the insertion, and all produced some progeny in which the insertion was detected. Overall, 60% of the parents contained the insertion compared to 3% of single L1 offspring and 75% of single adult offspring (Table 4.3). We examined insertion into the *unc-54* hotspot in the wild-type strain Bergerac. As with TW332, we screened Bergerac animals for insertions in parental hermaphrodites and their larval and adult offspring. The results are similar to those obtained for TW332. Insertion into the hotspot was detected in 40% of the parent worms, 2% of the L1 offspring, and 66% of the adult progeny (Table 4.3). These observations are consistent with the insertions occurring in somatic tissues during development. Collectively, these results indicate that most or all inserts we detect are in somatic cells and that these events occur almost exclusively in post-embryonic development.

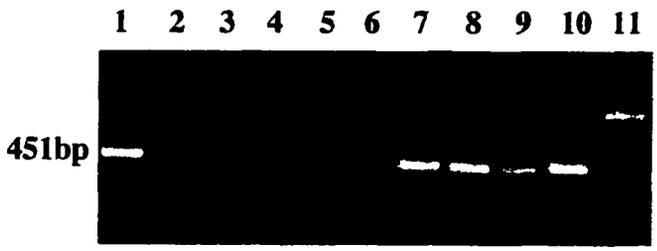


Figure 4.4 PCR products from single animals amplified with nested primers JC58 and JC67. Lane 1 contains a 451bp product amplified from a single TW332 "parent". Lanes 2-6 are products from single L1 offspring and lanes 7-11 are from adult offspring.

Table 4.3 Summary of insertion frequencies in parental worms and their larval and adult offspring.

strain	# animals screened	# insertions into hotspot	frequency
332 parents	5	3	0.60
332 L1 progeny	40	1	0.03
332 adult progeny	40	30	0.75
EM1002 parents	5	2	0.40
EM1002 L1 progeny	50	1	0.02
EM1002 adult progeny	50	33	0.66

Tc1 insertions into *unc-54* are detected in adult worms lacking a germline

While the experiments described above strongly suggest that frequent Tc1 insertions into *unc-54* are somatic, the results could also be explained if the insertions occur in germ tissue that is not represented in the next generation. *C. elegans* adults can produce more gametes than they do progeny (Wood, 1988). TW332 hermaphrodites have brood sizes of approximately 30, compared to 300 for N2 adults. TW332 may produce a far greater number of germ nuclei than progeny. Since PCR amplification can occur from template molecules from germ nuclei or somatic cells, it is possible that the frequent insertion into *unc-54* occurs in germ nuclei which are not inherited.

As a more definitive test of the idea that these insertions are somatic, we examined strains without a germline for Tc1 insertions into *unc-54*. Strains containing the *glp-4(bn2)* allele produce normal numbers of germ nuclei when raised at the permissive temperature (16°C) and very few germ-nuclei when raised at the restrictive temperature (25°C). Beanan and Strome (1992) report approximately 12 germ nuclei in young adults homozygous for the *glp-4(bn2)* allele raised at 25°C in contrast to the 700-1000 produced by wild-type adults. The *glp-4* mutation was isolated in a Bristol genetic background so we crossed the *glp-4* strain by TW332 and EM1002 and examined Tc1 insertion in progeny raised at 16°C and 25°C.

A high frequency of insertion was detected in F2 lines raised at 16°C. For TW332 and Bergerac derived strains, insertions are detected among the single animals screened as well as in the pools of ten worms (Table 4.4). Insertion into the hotspot is also frequent in worms raised at 25°C. The *glp-4(bn2)* strains show reduced levels of Tc1 insertion compared to the parent strains TW332 and Bergerac. Although less abundant in the *glp-4(bn2)* strains, the insertions appear to be in somatic tissues since the frequency of Tc1

Table 4.4 Summary of somatic insertion frequencies in *glp-4(bn2)* strains. The number 16 or 25 following a strain name refers to the temperature at which the animals were raised. Samples prepared from pools of ten animals are followed by the abbreviation pop10.

strain	#animals or populations screened	# insertions into hotspot	frequency
332 X *glp-4* 16	25	1	0.04
332 X *glp-4* 25	25	1	0.04
EM1002 X *glp-4* 16	25	1	0.04
EM1002 X *glp-4* 25	25	1	0.04
332 X *glp-4* 16 pop10	30	30	1.00
332 X *glp-4* 25 pop10	30	30	1.00
EM1002 X *glp-4* 16 pop10	16	10	0.63
EM1002 X *glp-4* 25 pop10	20	11	0.55

insertion into *unc-54* is approximately the same between worms depleted in germ nuclei and those that produce a normal germline.

glp-4(bn2) animals produce significantly fewer germ nulcei than wild-type animals but still produce an increasing number of germ nuclei as the animals age, although at a rate much slower than wild-type (Beanan and Strome, 1992). Additionally, construction of the *glp-4* strains results in a change in the TW332 genetic background and a significantly lower level of Tc1 insertion than TW332. To unambiguously rule out the possibility that the frequent insertions we detect in TW332 occur in the germline, we prepared animals which lack all germ tissue and screened their DNA for *unc-54* insertions.

Two cells, Z2 and Z3 give rise to the entire *C. elegans* germline. To generate animals completely without germline, I ablated Z2 and Z3 cells with a laser microbeam in early L1 larvae from strain TW332. Ablated animals were allowed to mature, giving rise to adults completely lacking germ-line tissue. DNA was prepared from 51 single adults lacking germ tissue as well as 58 adults which were not ablated but were collected from the same plate of TW332 as the ablated animals. Each template was screened for Tc1 insertions using PCR. Frequent Tc1 insertion is detected among ablated and unablated animals. Figure 4.5 shows typical PCR products amplified from single TW332 adults completely lacking germ tissue. The frequency of insertion into the *unc-54* hotspot is approximately the same between ablated and non-ablated TW332 adults (Table 4.5). Bands in addition to

Table 4.5 Summary of somatic insertion frequencies in TW332 animals with germlines ablated and without ablation.

strain and treatment	# animals screened	# insertions into hotspot	frequency
TW332 Z2&Z3 ablated	51	36	0.71
TW332 not ablated	58	42	0.72

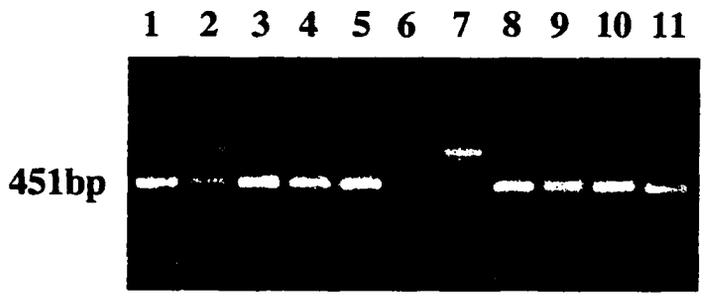


Figure 4.5 PCR products amplified from single adult hermaphrodites. Lane 1 contains a 451 bp product amplified from strain TR1299. Lanes 2-11 contain products amplified from TW332 worms which completely lack a germline due to laser ablation of germ-line precursor cells early in development.

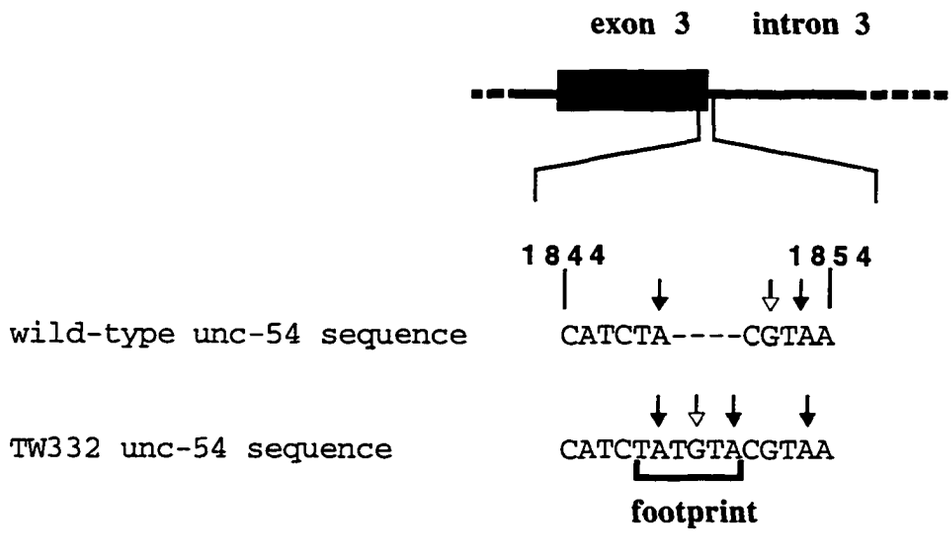


Figure 4.6 The diagram shows the exon3/intron3 boundary in the *unc-54* gene. The sequence of this region from wild-type as well as TW332 animals is shown below the map. Numbers above the sequence correspond to positions in the *unc-54* gene (Karn et al,). Shaded arrows denote TA dinucleotides which are frequent targets for somatic insertion of Tc1. TW332 contains a four base insertion compared to wild-type animals. The insertion is contained in the region labeled footprint. Because the footprint sequence both begins and ends with the dinucleotide TA we cannot determine if the 4 bp insertion is TATG or TGTA. The presence of the footprint alters splicing of the third intron. The 5' splice donor sequences are indicated by unshaded arrows above the sequences. A splice site 4 bp upstream of the wild-type donor is used preferentially in TW332.

the hotspot insertion are also seen among ablated and non-ablated worms at approximately the frequency expected for TW332 (see above). The only explanation for detecting new insertion events in animals without a germline is that the insertions occur in somatic tissues.

The sequence of the *unc-54* hotspot varies between strains

I have detected frequent insertions into the hotspot in *unc-54* in a variety of strains. Sequencing of the site in *unc-54* where Tc1 inserts at high frequency revealed a polymorphism between strains. TW332 contains a four base insertion at the hotspot relative to Bergerac, DH424, TR403, and Bristol (Figure 4.6). The wild-type sequence TA is replaced with TATGTA yielding a 4 bp insertion in the *unc-54* third exon. This insertion is probably a footprint left behind when Tc1 excised from TR674. Footprints are often generated upon Tc1 excision (Ruan and Emmons, 1987; Kiff et al., 1988; Eide and Anderson, 1988), and TATGTA is the most common footprint observed for Tc1 excision from this site (Carr and Anderson 1995). This +4 bp footprint results in an apparent frameshift in translational reading frame. However, Carr and Anderson (1995) have shown that the TGTA excision footprint results in the creation of a new 5' splice site 4 bp upstream of the normal 5' splice site in the *unc-54* third intron (Figure 4.6). The upstream splice site is used preferentially, removing the 4 bp Tc1 footprint from the mature mRNA. Altered splicing restores the translational reading frame of the transcript.

The TATGTA footprint in TW332 creates a new potential insertion site for Tc1 (Figure 4.6). Tc1 always inserts into the dinucleotide TA and wild-type *unc-54* contains two TAs within the interval 1848-1853. Neither of these sites is lost in TW332 and overall, an additional TA is gained. Sequencing of PCR products (see below) suggests that insertions occur at all of these TAs in TW332.

Sequences of insertion sites

To determine the precise location of Tc1 insertions in *unc-54* we directly sequenced PCR products amplified with primers JC67 and JC58. Twelve PCR products of approximately 450 bp amplified from single, adult, non-ablated TW332 hermaphrodites were sequenced. One outcome of directly sequencing PCR products is the possibility of sequencing multiple PCR products which comigrate on gels. Six sequences clearly indicate that the products are the result of a Tc1 insertion occurring at nucleotide position 1850 in the *unc-54* gene. Six additional products had sequences consistent with the presence of two or more Tc1 insertions at or near the hotspot in *unc-54*. PCR amplification of Tc1 insertions found within several nucleotides of each other generates products that comigrate on agarose gels and produces sequences with heterogeneity near the sites of insertion. These sequences are probably derived from single animals which contain an insertion at position 1850 as well as insertion at a nearby TA dinucleotide (of which there are 15 within the 100 bp of sequence flanking position 1850). Multiple insertions within an individual must occur in separate copies of *unc-54* since several insertions into the same copy would result in detection of a PCR product from only the insertional junction closest to the *unc-54* primer site.

The sequence of the hotspot region of *unc-54* in TW332 reveals that there are 3 potential Tc1 insertion sites within a 10 bp segment of the gene (Figure 4.6). The first base of Tc1 is C. The sequence of the gene and transposon junction when Tc1 inserts at position 1850, (after the first of the 3 TAs) is CTAC. When Tc1 inserts into the second TA (in the "footprint") or the third TA (at position 1854 in wild-type *unc-54*) the sequence created at the junction is GTAC, an *RsaI* restriction site. *RsaI* digestion of PCR products amplified from single TW332 animals reveals that some products, which migrate as ~450 bp products, are cut with *RsaI*. Products from 16 TW332 worms were cut with *RsaI*. Seven

did not cut at all and presumably arise from animals containing an insertion at position 1850 only. Six products cut partially producing one product consistent with insertion at site 1850 and a second product representing insertion into the footprint or the third TA. Three products cut completely, indicating that they were derived from templates containing an insertion in the footprint or the third TA only. This indicates that all three nucleotide positions in this region of *unc-54* are hotspots for somatic Tc1 insertion. Insertion into all three sites is detectable among single animals, although the frequency of insertion seems to be highest into the first TA. Eleven out of 16 PCR products are dervived from Tc1 insertions into the first TA and 6 out of 16 are from insertions into the other sites.

Sequences of PCR products amplified from single ablated TW332 animals are similar to those from non-ablated animals. Out of 20 ~450 bp PCR products sequenced, 10 are clearly from insertions at position 1850, 3 insertion sequences are in the footprint, one is at the third TA and 6 sequences are from multiple templates.

Ten PCR products of size greater than or less than 450 bp were sequenced. All represented Tc1 insertions in *unc-54*. All insertions occurred at TA dinucleotides. Insertion sites included positions 1543, 1699, 2014, 2140, 2143, and 2796 in *unc-54*. Five of the ten sequences were insertions at position 2014 in the third intron. The additional bands sequenced do not represent a random sample of larger and smaller PCR products, and the repetition of certain insertion site sequences is not necessarily representative of the frequency of insertion at that site. Bands of sizes other than 450 bp are observed frequently in single animals and bands of a particular size class are sometimes observed in several individuals (e.g. the 615 bp product generated by insertion at position 2014). This region of the *unc-54* gene appears to contain many potential targets for somatic insertion of Tc1.

Tc1 inserts frequently into another region of *unc-54*

Most of the somatic insertions we detect in the hotspot region of *unc-54* are into the same site where Eide and Anderson (1988) isolated 7 out of 11 spontaneous Tc1 induced *unc-54* germ-line mutations in Bergerac. To determine if this site in *unc-54* is exceptional, we screened another region of *unc-54* with nested primer set JC73 and JC74 that anneal in exon 5 (Table4.1; Figure 4.1). PCR performed on templates from 50 single TW332 adults and ten pools of ten adults generated several different products. A product of approximately 900bp was detected in three out of fifty individuals and in four out of ten pools of worms. Sequencing of the 900bp product from one individual revealed a Tc1 insertion at position 3715 in exon 6 of *unc-54.* This same site is represented once among the 11 germ-line insertions characterized by Eide and Anderson (1988). This suggests that sites that are frequent targets for germ-line insertion of Tc1 are hotspots for somatic insertion of Tc1 as well.

Another *C. elegans* gene, *src-1*, contains hotspots for somatic insertion of Tc1

To examine whether the *unc-54* gene is unusual in containing hotspots for somatic insertion of Tc1, I screened for insertions of Tc1 in another C. elegans gene, *src-1*. This gene encodes a presumed *C. elegans* homologue of the vertebrate oncogene *src* (Thacker, personal communication). Using primers JC61 and JC62 (see Table 4.1), two primers specific for an exon in *src-1*, and the Tc1 primers described above, we amplified and sequenced products from small populations of strain TW332. Two sites are identified as hotspots within this region of *src-1*, although neither is hit as frequently as the sites in *unc-54*. Insertion into each of these sites was detected in 8 out of 10 populations of 100 TW332 worms screened. This demonstrates that sites in other genes are frequent targets for Tc1 insertion although at levels less than that observed for *unc-54*. I did not screen for *src-1* insertions in animals lacking a germline and therefore cannot be sure that they occur

primarily in somatic cells. However, repeated attempts to isolate animals homozygous for these two frequent *src-1* insertions using a sib-selection protocol were unsuccessful suggesting that they are somatic (data not shown).

Discussion and Conclusions:

Tc1 inserts at high frequency in somatic cells:

We investigated the ability of Tc1 to insert in somatic cells. In the *mut-2* strain TW332, almost every animal contains an insertion at the hotspot in *unc-54*. Many individuals contain insertions into other sites in the *unc-54* gene and into other genes. The high frequency of Tc1 insertion into the unc-54 gene is not confined to the *mut-2* genetic background. Insertion is frequent in most wild-type strains but not in the common laboratory strain Bristol. The frequent Tc1 insertions must be confined to somatic tissues since they are detected in adult worms lacking a germline. In addition to tissue-specific regulation of transposition, somatic insertion of Tc1 may be developmentally regulated since insertions are rarely detected in L1 larvae but are abundant in adults.

Somatic insertion of Tc1 occurs at very high frequency and may represent a significant source of spontaneous mutation in somatic tissue. In the strain TW332, at least 71% of single animals contain an insertion at a single site in the *unc-54* gene. This value may be an underestimate since only insertions in one orientation are considered. Additionally, somatic insertions present in one or a few cells may not be detected if insertion-containing templates are damaged or lost during DNA preparation and handling. I observed 36 out 51 ablated TW332 animals containing an insertion at the hotspot. Assuming that the number of insertions per worm is Poisson distributed, the probability of observing zero insertions in a sample is $p(X=0)=e^{-\lambda}$, where λ is the rate of insertion. Estimating the $p(X=0)$ as 1-

p(hotspot insertion is amplified from a single ablated worm)=1-(36/51)=0.29, I estimate λ to be 1.2 insertions per worm. *C. elegans* has 1918 somatic genomes. Therefore the expected probability that a single copy of *unc-54* contains an insertion is 1.2/1918=6.2 X 10^{-4}. If each of the 13,000 or so *C. elegans* genes contains a hotspot like the one observed in *unc-54*, then we would expect to find approximately eight genes containing an insertion in every copy of the genome or about 15,500 new somatic insertions in each animal. As suggested by my observation of frequent insertion of Tc1 in both orientations into the hotspot, these estimates of the number of somatic insertions are probably underestimates. Because the PCR based screen limits our ability to detect insertions occurring more than ~1.5kb from the *unc-54* primers selected, it is possible that *unc-54* contains additional, as yet undetected, hotspots. Even if the hotspot in *unc-54* is exceptional, results for a second site in *unc-54* as well as a site in *src-1* indicate that other sites experience insertion at frequencies within one or two orders of magnitude of that observed for the *unc-54* hotspot. Somatic transposition may represent a significant mutational load for an individual.

Regulation of somatic Tc1 activity:

Somatic mutations which occur very early in development have the potential to rise to high frequency within a single animal as a result of cell proliferation. If these mutations are deleterious, we might predict that natural selection would favor the evolution of mechanisms which restrict the somatic movement of transposons to later stages of development. The observation that L1 larval worms have approximately forty-fold less frequent insertion of Tc1 into the *unc-54* hotspot as compared to adult worms yet contain about 2-fold fewer cells suggests that somatic insertion is actively suppressed during early stages of development. If somatic insertion was equally likely during all cell divisions we would expect only about a two-fold difference in frequency between adults and L1 larvae.

The crosses performed with *glp-4(bn2)* demonstrate that somatic transposition is heritable. To create strains that exhibit high levels of somatic insertion and contain a mutation (*glp-4(bn2)*) reducing the number of germ nuclei, we crossed a strain with very low levels of somatic transposition, which was isolated in a Bristol genetic background, to strains TW332 and Bergerac that show very high levels of somatic insertion. Since the resulting strains show levels of insertion higher than Bristol, somatic activity must be inherited. However, the mode of inheritance appears to be complex. The *glp-4(bn2)* derived strains had levels of somatic insertion intermediate between those of the parent strains suggesting that inheritance of the somatic mutator phenotype is not simply the result of inheritance of a single gene. It is possible that regulation of somatic transposon activity is a polygenic trait. It may be polygenic in the sense that several genes are responsible for regulating activity or alternatively, that regulation depends on the number of copies of an element in the genome. The results of the crosses do not distinguish between these two potential explanations since additional copies of Tc1 could be inherited in addition to somatic mutator loci. The two strains derived in the *glp-4(bn2)* crosses are expected to have intermediate number of copies of Tc1 and an intermediate frequency of somatic insertion if somatic activity is copy number dependent. However, if the trait is polygenic, and the high levels of somatic insertion are due to the additive effects of alleles at several loci, we might also expect to see reduced levels of activity in strains derived from our crosses.

It is known that transposition and excision of Tc1 in *C. elegans* are regulated in a strain- and tissue-specific manner (Moerman and Waterston, 1989). Although multiple Tc1 sequences are found in the genome of every *C. elegans* isolate, activity of Tc1 is restricted to certain genetic backgrounds. Tc1 elements insert and excise at low or undetectable frequencies in the germlines of Bristol (Moerman and Waterston, 1984; Emmons and Yesner, 1984) and DH424 (Eide and Anderson, 1985) isolates. Tc1 insertion is the major

cause of spontaneous germline mutation in Bergerac (Moerman and Waterston, 1984; Eide and Anderson, 1985; Moerman et al., 1986) and TR403 isolates (Phil Anderson, personal communication). TW332 contains the *mut-2(r459)* mutator allele which leads to levels of germ-line Tc1 insertion fifty-fold higher than that of Bergerac (Collins et al., 1987). Germline excision of Tc1 is observed only in strains where the element also actively inserts in the germline. Somatic excision of Tc1, on the other hand, occurs at frequencies several orders of magnitude higher than in the germline and shows little variation in different genetic backgrounds. Somatic excision frequencies for several Tc1 alleles are no more than a tenfold lower in Bristol than in Bergerac (Harris and Rose, 1986). In TW332, where germ-line excision frequencies are fifty-fold higher than Bergerac, somatic excision frequencies do not appear elevated (Collins et al., 1987). Overall, somatic excision frequencies appear very similar between different strains.

Somatic insertion frequencies may be more sensitive to genetic background than somatic excision. One obvious difference between somatic insertion and excision is apparent in Bristol. Levels of somatic excision are comparable between Bristol and other strains whereas somatic insertion is rare in Bristol. This difference in somatic insertion frequencies between strains might arise as a result of variation in Tc1 copy number. Bristol contains about ten to twenty-fold fewer copies of Tc1 than any other strain tested and has the lowest level of somatic insertion. Insertion is most frequent in strains which are expected to have the highest copy number for Tc1 (TW332 and Bergerac).

There are at least two plausible mechanisms that could lead to copy number dependent somatic insertion frequencies. Either an element encoded factor or excision products of the element could be involved in somatic insertion. It is known that overexpression of a construct containing Tc1 coding sequence results in an increase in the frequency of insertion into the *gpa-2* gene (Vos et al., 1993). This suggests that Tc1 transposase is a limiting factor in the transposition process. It is possible that strains with a higher copy

number of Tc1 produce more transposase and hence, higher frequencies of somatic insertion. Alternatively, the availability of excision products may affect rates of somatic insertion. Extrachromosomal copies of Tc1 have been identified in *C. elegans* and may represent intermediates for insertion (Ruan and Emmons, 1984; Radice and Emmons, 1993). If excision products are a limiting intermediate for transposon insertion, strains with high levels of excision should show high levels of insertion. The total pool of excision products in a cell should be a function of the number of elements capable of excision as well as the frequency with which they excise. Although frequencies of somatic excision for an individual Tc1 element are comparable between strains, the total number of available excision products may vary as a function of element copy number.

Insertion of Tc1 in the germ-line does not appear to be determined entirely by copy number. Transposition of Tc1 is undetectable in the genomes of both N2 and DH424 (Eide and Anderson, 1985). DH424 has about ten times as many Tc1 elements as Bristol, yet no detectable insertion in its germline. The detection of high levels of somatic Tc1 insertion in DH424 but not in Bristol suggests that the frequencies of somatic insertion may not always be correlated with frequencies of germ-line insertion. The somatic and germ cell lineages in *C. elegans* consist of approximately the same number of cell divisions and generate roughly equal numbers of cells (Hirsh et al., 1976; Sulston and Horvitz, 1977; Kimble and Hirsh, 1979; Sulston et al., 1983). If transposon insertion was simply correlated with a cell-cycle associated event such as DNA replication we might expect to observe similar frequencies of insertion in both cell types. Since Tc1 insertion is orders of magnitude more frequent in the soma than in the germ-line, some additional explanation for the difference is required.

It is possible that differences arise because of a fundamental difference between the germ and soma. A potential explanation is that factors required for transposition are regulated by tissue-specific regulatory molecules. Alternatively, it is possible that some sites in the genome are more acccessible for insertion in somatic cells. Differences in the

accessability of sites between the germ-line and soma might arise from differences in chromatin structure or transciptional activity. However, at least some target sites are used in both the germline and the soma suggesting that any differences in gene structure and expression do not dramatically alter the pattern of insertion. Further study is required to sort out the mechanisms responsible for regulation of Tc1 activity in the germline and soma.

What makes a hotspot hot?

Eide and Anderson (1985) isolated 11 spontaneous Tc1-induced germ-line mutants in unc-54. Remarkably, 7 out of 11 insertions occurred at a single site in the gene (Eide and Anderson 1988). The somatic hotspot identified in our study is at the same site as the germ-line hotspot. We detected insertion of Tc1 into another site in *unc-54* in 3 out of 50 single animals. This site was also identified once in Eide and Anderson's (1988) collection of germ-line insertions into the *unc-54* gene. Although the regulation of Tc1 activity is tissue specific, the distibution of sites experiencing insertion may be similar in the different tissue types. This suggests that the machinery involved in Tc1 target site selection and element insertion are common to both tissue types.

The finding that Tc1 inserts at high frequency into the same site in both somatic and germ cells suggests that something about this region of the *unc-54* gene makes it a preferred target for Tc1 insertion. Primary, secondary or higher order structure (e.g. chromatin or DNA associated factors involved in transcription) of the target sequence may contribute in the definition of a hotspot. All known Tc1 insertions occur at the dinucleotide TA. Eide and Anderson (1988) proposed a consensus sequence for Tc1 insertion GA G/T A/G TA T/C G/C T. The sequence of the *unc-54* hotspot matches the consensus at 7 out of 9 positions. A polymorphism between TW332 and Bergerac alters the sequences flanking one side of the target site yet this site is a hotspot in both strains. In TW332, the

region of *unc-54* a few bases downstream of the hotspot contains two TA dinucleotides that are also frequent targets for insertion. The sequences flanking these other two TA dinucleotides differ from each other and from the sequence of the hotspot, but also match the consensus at 7 out of 9 positions. However, other sites in *unc-54* which are as good a match to the consensus as the hotspot do not appear to be frequent targets for insertion. It is not clear if these three TAs are frequent sites for insertion because they are all flanked by sequences preferred for Tc1 insertion or because of some other feature found in this region of *unc-54*. For Tc1 insertions in the *gpa-2* gene, van Luenen and Plasterk (1994) report only a weak correlation between the number of insertions at a particular TA dinucleotide and the match of the insertion site with the consensus sequence. Additionally, they found that hotspots for insertion were not clustered. Tc1 insertion is obviously constrained by target sequence, but it seems unlikely that the primary sequence of a region is the sole determinant of insertion site preference.

Secondary structures, such as bends or kinks in the DNA, could play a role in determination of insertion site preference. Inverted and direct repeated sequences may be useful in demarcating regions of secondary structure. However, these structures are common features of the *C. elegans* genome and so-far, no particular secondary structures are consistently associated with sites of Tc1 insertion (van Luenen and Plasterk, 1994). It is also possible that insertion site preference is determined by higher order structures of the target region. A precedent for such a situation is illustrated by the integration retroviral elements, where target site selection is affected by the transcriptional state of the target region and the distribution of nucleosomes on the DNA (Pryciak and Varmus 1992). Differences in chromatin structure between different regions of a gene or in the spatial distribution of other DNA-associated factors DNA might lead to enhancement or repression of insertion into different regions of a gene. The distribution of nucleosomes and DNA associated factors is unknown for the *unc-54* locus. Therefore, the high frequency of

insertion into some sites in *unc-54* may reflect some as yet unidentified structure in the *unc-54* locus.

Somatic transposition and reverse genetics:

Understanding how transposon activity is regulated in different cell types and how target sites are selected is important for the improvement of transposons as tools for reverse genetic approaches in *C. elegans*. As the *C. elegans* genome project approaches completion and the sequences of all 13,000 *C. elegans* genes are identified, efforts will be focused on determination of the biological function of each gene. At present, in the *C. elegans* research community, transposons provide the only means for altering gene sequences in a targeted fashion *in vivo*. I was using a sib-selection PCR approach (Ballinger and Benzer, 1989; Kaiser and Goodwin, 1990; Rushforth et al., 1993) to address the phenotypic consequences of germ-line Tc1 insertions into *unc-54*. Germ-line insertions proved difficult to isolate for *unc-54* because of the high levels of somatic insertion into this locus. Somatic insertion may be a major obstacle for isolating germline insertions in other loci as well.

PCR-based methods for detecting Tc1 insertions into *C. elegans* genes are widely used. Discrimination between PCR products generated from animals containing germ-line insertion and those where insertion occurs in somatic cells allows for more efficient isolation of germ-line mutants. Efforts to distinguish germ-line insertions from somatic insertions among PCR products amplified from a frozen mutant bank of *C. elegans* (Zwaal et al., 1993) has lead to the successful isolation of many germ-line insertions. This method relies on "semi-quantitative" PCR of DNA prepared from fairly small populations of worms. An insertion occurring in the germ-line of an individual and present in a portion its progeny is expected to generate a greater proportion of insertion-containing template molecules than is expected for infrequent somatic insertions. Conditions for PCR can be

adjusted to detect products only from abundant templates, thus eliminating detection of some somatic insertion products. Frequent somatic insertions into a particular site, like the hotspot in *unc-54*, may lead to false positives even when PCR is quantitative. Successful isolation of Tc1 insertions may be due to fortuitous selection of gene regions that are not frequently targets for somatic insertion. However, my data for *unc-54* suggests that hotspots are the same for germ-line and somatic insertion. Additional methods may be required to isolate germ-line insertions at some sites.

This study of somatic insertions into *unc-54* suggests several potential improvements for detection of element insertions. Screening of larger populations of animals by PCR followed by enrichment for germ-line mutants by sib-selection (Rushforth et al. 1993) is more likely to suffer from false positives arising from frequent somatic insertion. Screening smaller populations may reduce the proportion of somatic insertion templates relative to germline insertion templates and increase the likelihood of discriminating between them. Quantitative PCR should be useful in distinguishing between somatic and germ-line events. Additionally, problems associated with somatic insertion might be alleviated by screening for insertions in animals before extensive somatic insertion occurs. Somatic insertion into the *unc-54* hotspot is rare among L1 larvae and suggests that screening for germ-line insertions among L1 larvae or embryos might reduce detection of somatic insertions. Choosing a strain for reverse genetic approaches with Tc1 is critical. Tc1 must be active in the germline of the strain and preferably not move in somatic cells. We find that Tc1 inserts at high frequency in somatic cells of all strains except Bristol. This presents a problem since Tc1 is not active in the germline of Bristol animals. Ideally, a strain would be identified with a high level of germline activity (like TW332 and Bergerac) and a low level of somatic activity (like Bristol). Finally, somatic insertions could be avoided by using a transposon that does not move in somatic tissues. Preliminary results indicate that Tc5 elements move less frequently in somatic cells than Tc1 (Tc5

element excision is not detectable on Southern Blots, Collins, 1994) and may represent a better choice for reverse genetic approaches. Regardless of which element is chosen and which method is used to isolate new element insertions, subsequent analyses of gene function often requires additional manipulation of transposon-containing mutant alleles of a gene. Germ-line insertions of Tc1 may be used to isolate deletion derivatives or gene replacement products of the original Tc1 allele. These techniques also rely on screening populations of animals with PCR followed by enrichment by sib-selection and can be confounded by events occurring in somatic tissue. Further characterization of the factors regulating transposon activity will lead to improvements in these techniques.

Evolutionary significance of somatic transposition

Somatic mutation is seldom considered of great importance in evolution because the variation generated in somatic cells is not heritable, except in the sense that somatic mutations may be passed on within somatic cell lineages. However, mutations occurring in somatic tissue are not neccessarily without consequence. In a recent paper Orr (1995) proposes that the deleterious consequences associated with somatic mutation may have provided the conditions necessary for the evolution of diploidy. Since the likelihood of homozygosity of deleterious recessive alleles in somatic cells is reduced in diploids, they may be at a selective advantage over haploids. Insertion of transposons may be a major source of spontaneous mutation in somatic cells. This could lead to the evolution of mechanisms that reduce the deleterious consequences of insertion. Exactly how deleterious somatic insertions are and what mechanisms exist to control this behavior is unclear.

Like Tc1, mariner elements in *Drosophila* display high levels of somatic activity in some strains. Regulation of this activity results from the presence of a single dominant genetic factor, Mos (Bryan and Hartl, 1988), which is itself a Mariner element (Medhora et al., 1991). Strains where mariner elements actively move in the soma have reduced lifespans

compared to strains lacking somatic activity (Woodruff, 1993; Nikitin and Woodruff, 1995). The activity of P elements in *Drosophila melanogaster* is normally restricted to the germline because of differential splicing of the P element message in the germline and soma (Laski et al., 1986). However, P element constructs lacking the regulatory third intron, produce active transposase in somatic cells and a resulting increase in transposition in the soma. This activity also resulted in a shortening of life span (Driver and McKechnie, 1992). Since somatic transposition could affect a large number of different loci, we suspect that it may affect other components of fitness as well. There are likely to be many genes essential for somatic cell viability, a subset of which (e.g. oncogenes) will significantly affect fitness when mutated. Variation at loci which alter somatic mutation rates may play an important role in evolution.

Charlesworth and Langley (1986) suggest that in organisms where the germline and soma are developmentally distinct, transposition in somatic cells confers no selective advantage to transposable elements, because there is no possibility of transmission to the next generation. In fact, it is likely to be disadvantageous since somatic mutations may reduce the liklihood that an individual reproduces. So, selection is expected to favor elements that do not transpose in somatic cells. In organisms where the distinction between the germline and soma is less clear, such as in plants, somatic transposition may lead to transmission to the next generation and may be advantageous for a transposon. In *C. elegans* the distinction between the germline and soma is apparent as early as the 4 cell stage (Wood, 1988). The observation of high frequencies of somatic transposition for Tc1 is somewhat surprising. There are at least 4 possible explanations for this activity. Transposition in somatic cells might be favored if a mechanism existed for the introduction of somatic insertion products into the germline. However it is unlikely that such a mechanism exists in *C. elegans*. Second, somatic activity might be selectively advantageous for the element if it leads to an increase in the probability that an element

experiences horizontal transmission. Horizontal transmission of transposons is often invoked as a mechanism by which transposons persist over long evolutionary periods of time (Capy et al., 1994). However, horizontal transfer of elements is expected to be a rare event making it unlikely that selection could maintain somatic transposition in anticipation of the occasional benefits conferred by horizontal transfer. Third, somatic transposition may be slightly disadvantageous but persist because the factors necessary for transposition in the germline are not strictly confined to this tissue type. Finally, somatic transposition coupled with high levels of somatic excision (as observed for Tc1) may render activity in the soma selectively neutral.

LIST OF REFERENCES

Ajoika, J.W. and D.L. Hartl. 1989. Population dynamics of transposable elements. Pages 939-958. In D. Berg and M. Howe, (eds). Mobile DNA.

Altschul, S., W. Gish, W. Miller, E. Myers and D. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-10.

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

Babity, J., T. Starr and A. Rose. 1990. Tc1 transposition and mutator activity in a Bristol strain of C. elegans. Mol. Gen. Genet. 222:65-70.

Baker, T.A. and K. Mizuuchi. 1992. DNA-promoted assembly of the active tetramer of the Mu transposase. Genes Dev. 6:2221-2232.

Ballinger, D.G. and S. Benzer. 1989. Targeted gene mutations in Drosophila. Proc. Natl.Acad. Sci. USA 86:9402-9406.

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

Beanan, M.J. and S. Strome. 1992. Characterization of a germ-line proliferation mutation in C. elegans. Development 116:755-766.

Benian, G., S. L'Hernault and M. Morris. 1993. Additional sequence complexity in the muscle gene, unc-22, and its encoded product, Twitchin, of C. elegans. Genetics 134:1097-1104.

Berg, D. and M. Howe (ed). 1989. Mobile DNA. American Society of Microbiology, Washington, D.C.

Black, D., M. Jackson, M. Kidwell and G. Rubin. 1987. KP elements repress P-induced hybrid dysgenesis in D. melanogaster using a novel and general method. Cell 25:693-704.

Boeke, J. Eichinger, D. and G. Fink. 1989. Regulation of yeast Ty element transposition. Pages 169-180. In M.E. Lambert, J.F. McDonald and I.B. Weinstein, (eds). Eukaryotic transposable elements as mutagenic agents. Cold Spring Harbor Press, Cold Spring Harbor, N.Y.

Britten, R.J. and E.H. Davidson. 1969. Gene regulation for higher cells: a theory. Science 165:349-357.

Bryan, G., D. Garza and D. Hartl. 1990. Insertion and excision of the transposable element mariner in Drosophila. Genetics 125:103-114.

Bryan, G., J. Jacobson and D. Hartl. 1987. Heritable somatic excision of a Drosophila transposon. Science 235:1636-1638.

Bryan, G. and D. Hartl. 1988. Maternally inherited transposon excision in D. simulans. Science 240:215-217.

Calui, B., T. Hong, S. Findley, W. Gelbart. 1991. Evidence for a common evolutionary origin of inverted repeat transposons in Drosophila and plants: Hobo, Activator and Tam3. Cell 66:465-471.

Campbell, A. 1983. Transposons and their evolutionary significance. Pages 258-279. In M. Nei and R.K. Koehn (eds). Evolution of genes and proteins. Sinauer Associates, Sunderland, Massachusetts.

Capy, P. D. Anxolabehere and T. Langin. 1994. The strange phylogenies of transposable elements: are horizontal transfers the only explanation? Trends Genet. 10(1):7-11.

Chao, L. and S. McBrown. 1985. Evolution of transposable elements: An IS10 inversion increases fitness in E. coli. Mol. Biol. Evol. 2:359-369.

Charlesworth, B. and C. Langley. 1986. The evolution of self-regulated transposition of transposable elements. Genetics 359-383.

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

Charlesworth, B., A. Lapid and D. Canada. 1992. The distribution of transposable elements within and between chromosomes in a population of Drosophila melanogaster. I. Element Frequencies and Distribution. Genetical Research 60:103-114.

Charlesworth, B., A. Lapid and D. Canada. 1992. The distribution of transposable elements within and between chromosomes in a population of D. melanogaster II. Inferences on the nature of selection against elements. Genetical Research 42:1-27.

Collins, J., E. Forbes and P. Anderson. 1989. The Tc3 Family of Transposable Genetic Elements in C. elegans. Genetics 121:47-55.

Collins, J., B. Sari and P Anderson. 1987. Activation of a transposon in the germline but not the soma of C. elegans. Nature 328:726-728.

Collins, J. and P. Anderson. 1994. The Tc5 family of transposable elements in C. elegans. Genetics 137:771-781.

Coulson, A., J. Sulston, S. Brenner and J. Karn. 1986. Toward a physical map of the genome of the nematode C. elegans. Proc. Natl. Acad. Sci. USA 83:7821-7825.

Cresse, A.D., S.H. Hulbert, W.E. Brown, J.R. Lucas and J.L. Bennetzen. 1995. Mu1-related transposable elements of maize preferentially insert into low copy number DNA. Genetics 140:315-324.

Devereux, J., P. Haeberli and O. Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12(1):387-395.

Dibb, N. J., I. N. Maruyama, M. Krause and J. Karn. 1989. Sequence analysis of the complete Caenorhabditis elegans myosin heavy chain gene family. J. Mol. Biol. 205: 603-613.

Doak, T.G., F.P. Doerder, C.L. Jahn and G. Herrick. 1994. A proposed superfamily of transposase genes: Transposon-like elements in ciliated protozoa and a common "D35E" motif. Proc. Natl. Acad. Sci. 91:942-946.

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

Dooner, H. and A. Belachew. 1989. Transposition pattern of the maize element Ac from the bz-m2(Ac) allele. Genetics 122:447-457.

Doseff, A., R. Martienssen and V. Sundaresan. 1991. Somatic excision of the Mu1 transposable element of maize. Nucleic Acids Res. 19(3):579-584.

Dreyfus, D. and S. Emmons. 1991. A transposon-related palindromic repetitive sequence from C. elegans. Nucleic Acids Res. 19(8):1871-1877.

Egilmez, N.K., R.H. Ebert, R.J. Shmookler Reis. 1995. Strain evolution in C. elegans: Transposable elements as markers of interstrain evolutionary history. J. Mol. Evol. 40:372-381.

Eide, D. and P. Anderson. 1985a. Transposition of Tc1 in the nematode C. elegans. Proc. Natl. Acad. Sci. USA 82:1756-1760.

Eide, D. and P. Anderson. 1985b. The gene structures of spontaneous mutations affecting a C. elegans myosin heavy chain gene. Genetics 109:67-79.

Eide, D. and P. Anderson. 1988. Insertion and Excision of C. elegans transposable element Tc1. Mol. and Cell. Biol. 8(2):737-746.

Emmons, S.W., M.R. Klass and D. Hirsh. 1979. Analysis of the constancy of DNA sequences during the development and evolution of the nematode C. elegans. Proc. Natl. Acad. Sci. USA 76:1333-1337.

Emmons, S.W., L. Yesner, K.S. Ruan and D. Katzenberg. 1983. Evidence for a transposon in C. elegans. Cell 32:55-65.

Emmons , S.W. and L. Yesner. 1984. High-Frequency Excision of Transposable element Tc1 in the Nematode C. elegans is Limited to Somatic Cells. Cell 36:599-605.

Emmons, S.W. S. Roberts and K.S. Ruan. 1986. Evidence in a nematode for regulation of transposon excision by tissue specific factors. Mol. Gen. Genet. 202:410-415.

Engels, W. R. 1989. P elements in Drosophila melanogaster. Pages 437-484. In M.E. Lambert, J.F. McDonald and I.B. Weinstein, (eds). Eukaryotic transposable elements as mutagenic agents. Cold Spring Harbor Press, Cold Spring Harbor, N.Y.

Engels, W.R., D.M. Jonson-Schlitz, W. B. Eggleston and J. Sved. 1990. High-frequency P element loss in Drosophila is homolog dependent. Cell 62: 515-25.

Errede, B., M. Company and C. Hutchinson. 1987. Ty1 sequence with enhancer and mating-type-dependent regulatory activities. Mol. Cell. Biol. 7:258-264.

Finnegan, D. J. 1989. Eukaryotic transposable elements and genome evolution. Trends Genet. 5(4):103-107.

Futuyma, D. 1986. Evolutionary biology. Sinauer Associates, Inc., Sunderland, Massachusetts.

Geyer, P., A. Chien, V. Corces and M. Green. 1991. Mutations in the su(s) gene affect RNA processing in D. melanogaster. Proc. Natl. Acad. Sci. USA 88:7116-7120.

Gloor, G.B., Nassif, N.A., Johnson-Schlitz, D.M., Preston, C.R., W.R. Engels. 1991. Targeted gene replacement in Drosophila via P element-induced gap repair. Science 253:110-117.

Gould, S. and R. Lewontin. 1979. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptionist programme. Proc. R. Soc. Lond. B205:581-598.

Green, M. 1989. Mobile DNA elements and spontaneous gene mutation. Pages 41-50 in Lambert, M., J. McDonald and I. Weinstein (eds). Eukaryotic transposable elements as mutagenic agents. Cold Spring Harbor Press, New York.

Hall, B. 1988. Adaptive evolution that requires multiple spontaneous mutations. I. Mutations involving an insertion sequence. Genetics 120:887-897.

Handler, A., S. Gomez and D. O'Brochta. 1993. Negative regulation of P element excision by the somatic product and terminal sequences of P in D. melanogaster. Mol. Gen. Genet. 237:145-151.

Harris, L.J. and A.M. Rose. 1986. Somatic excision of the transposable element Tc1 from the Bristol genome of C. elegans. Mol. and Cell. Biol. 6:1782-1786.

Harris, L.J. and A.M. Rose. 1989. Structural analysis of Tc1 elements in C. elegans var. Bristol (strain N2). Plasmid 22:10-21.

Hartl, D. and A. Clark. 1989. Principles of Population Genetics. Sinauer Associates, Inc., Sunderland, Massachusetts.

Hartl, D. and D. Dykhuizen. 1984. The population genetics of E. coli. Annu. Rev. Genet. 18:31-68.

Henikoff, S. 1992. Detection of C. elegans transposon homologs in diverse organisms. New Bio. 4:382-388.

Hirsh, D., D. Oppenheim and M. Klass. 1976. Development of the reproductive system of C. elegans. Dev. Biol. 49:200-219.

Horowitz, H. and C. Berg. 1995. Aberrant aplicing and transcription termination caused by P element insertion into the intron of a Drosophila gene. Genetics 139:327-335.

Houck, M., J. Clarke, K. Peterson and M. Kidwell. 1991. Possible horizontal transfer of drosophila genes by the mite Proctolaelaps regalis. Science 253:1125-1128.

Jacobson, J.W. and D. L. Hartl. 1985. Genetics 111:57

Ji, H., D.P. Moore, M.A. Blomberg, L.T. Braiterman, D.F. Voytas, G. Natsoulis & J.D. Boeke. 1993. Hotspots for unselected Ty1 transposition events on yeast chromosome III are near tRNA genes and LTR sequences. Cell 73: 1007-1018.

Junakovic, N., C. DiFranco, P. Barsanti and G. Palumbo. 1986. Transposition of copia-like nomadic elements can be induced by heat shock. J Mol Evol 24:89-93.

Kaiser, K. and S. Goodwin. 1990. "Site-selected" transposon mutagenesis of Drosophila. Proc. Natl. Acad. Sci. USA 87:1686-1690.

Karn J., S. Brenner and L. Barnett. 1983. Protein structural domains in the C. elegans unc-54 myosin heavy chain gene are not separated by introns. Proc. Natl. Acad. Sci USA 80:4253-4257.

Kaufman, P. and D. Rio. 1992. P element transposition in vitro proceeds by a cut-and-paste mechanism and uses GTP as a cofactor. Cell 69:27-39.

Kiff, J.E., D.G. Moerman, L.A. Schriefer, and R.H. Waterston. 1988. Transposon-induced deletions in unc-22 of C. elegans associated with almost normal gene activity. Nature 310:332-333.

Kim, H.-Y., J. Schiefelbein, V. Raboy, D. Furtek and O. Nelson. 1987. RNA splicing permits expression of a maize gene with a defective suppressor-mutator transposable element insertion in an exon. Proc. Natl. Acad. Sci. USA 84:5863-5867.

Kimble, J. and D. Hirsh. 1979. The post-embryonic cell lineages of the hermaphrodite and male gonads in C. elegans. Dev. Biol. 70:396-417.

Klein, A. S. and O. E. Nelson. 1983. Biochemical consequences of the insertion of a suppressor-mutator (Spm) receptor at the bronze-1 locus in maize. Proc Natl. Acad. Sci. USA 80:7591-7595.

Kobayashi, S., T. Hirano, M. Kakinuma and T. Uede. 1993. Transcriptional repression and differential splicing of FAS mRNA by early transposon (ETn) insertion in autoimmune LPR mice. Biochem. Biophys. Res. Commun. 191:617-624.

Kocher, T. D. and A.C. Wilson. 1991. DNA amplification by the polymerase chain reaction. Pages 187-209 in Essential Molecular Biology, Volume 2. T.A. Brown (ed) Oxford University Press.

Lambert, M., J. McDonald and I. Weinstein (eds). 1989. Eukaryotic transposable elements as mutagenic agents. Cold Spring Harbor Press, New York.

Laski, F., D. Rio and G. Rubin. 1986. Tissue specificity of Drosophila P element transposition is regulated at the level of mRNA splicing. Cell 44:7-19.

Levitt, A. and S.W. Emmons. 1989. The Tc2 transposon in C. elegans. Proc. Natl. Acad. Sci. USA 86:3232-3236.

Li, W. and J. Shaw. 1993. A variant Tc4 element in the nematode C. elegans could encode a novel protein. Nucleic Acids Res. 21:59-67.

Liao, L.W., B. Rosenzweig and D. Hirsh. 1983. Analysis of a transposable element in C. elegans. Proc. Natl. Acad. Sci. USA 80:3585-3589.

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

MacLeod, A.R., R.H. Waterston, R.M. Fishpool, and S. Brenner. 1977. Identification of the structural gene for a myosin heavy chain in C. elegans. J. Mol. Biol. 114:133-140.

McClintock, B., 1948. Mutable loci in maize. Carnegie Inst. Wash. Year Book 47:155-169.

McClintock, B., 1949. Mutable loci in maize. Carnegie Inst. Wash. Year Book 48:142-154.

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

McDonald, J.F.1990. Macroevolution and Retroviral elements. Bioscience 40(3):183-191.

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

McEntee, K. and V.A. Bradshaw. 1989. Effects of DNA damage on transcription and transposition of Ty retrotransposons of yeast. Pages 245-254. In M.E. Lambert, J.F. McDonald and I.B. Weinstein, (eds). Eukaryotic transposable elements as mutagenic agents. Cold Spring Harbor Press, Cold Spring Harbor, N.Y.

Medhora, M., K. Maruyama and D. Hartl. 1991. Molecular and functional anaysis of the mariner mutator element Mos1 in Drosophila. Genetics 128:311-318.

Menssen, A., S. Hohmann, W. Martin, P. Schnable, P. Peterson, H. Saedler and A. Gierl. 1990. The En/Spm transposable element contains splice sites at the termini generating a novel intron from a dSpm element in the A2 gene. EMBO J. 9(10):3051-3057.

Mills, M. 1993. Genetic and molecular analysis of the transposable elements Tc3 and Tc5 in C. elegans. Masters Thesis, University of New Hampshire, Durham, New Hampshire.

Miller, D. and Miller, L. 1982. A virus mutant with an insertion of a copia-like element. Nature 299:562-564.

Mizuuchi, K. 1983. In vitro transposition of bacteriophage Mu: a biochemical approach to a novel replication reaction. Cell 35:785-794.

Modi, R., L. Castilla, S. Puskas-Rozsa, R. Helling and J. Adams. 1992. Genetic changes accompanying increased fitness in evolving populations of E. coli. Genetics 130:241-249.

Moerman, D.G., G.M. Benian and R.H. Waterston. 1986. Moecular cloning of the muscle gene unc-22 by Tc1 transposon tagging. Proc. Natl. Acad. Sci. USA 83:2579-2583.

Moerman, D.G., G.M. Benian, R., J. Barnstead, L. Schriefer and R.H. Waterston. 1988. Identification and intracellular localization of the unc-22 gene product of C. elegans. Genes Dev. 2:93-105.

Moerman, D.G., J.K. Kiff and R.H. Waterston. 1991. Germline excision of the transposable element in C. elegans. Nucleic Acids Res. 19(20) 5669-5672.

Moerman, D.G. and R.H.Waterston. 1984. Spontaneous unstable unc-22 mutations in C. elegans variety Bergerac. Gnetics 108:859-877.

Moerman, D.G. and R.H.Waterston.1989. Mobile elements in C. elegans and other Nematodes. Pages 537-555. In Berg and Howe, (eds). Mobile DNA.

Montgomery, E. and C. Langley. 1983. Transposable elements in Mendelian populations. II. Distribution of three copia-like elements in a natural population of Drosophila melanogaster. Genetics:104:473-483.

Morawetz, C. 1987. Effect of irradiation and mutagenic chemicals on the generation of ADH2-constitutive mutants in yeast. Significance for the inducibility of Ty transposition. Mutat. Res. 177:53-60.

Mori, I., G. Benian, D. Moerman and R. Waterston. 1988. The transposon Tc1 of C. elegans recognizes specific target sequences for integration. Proc. Natl. Acad. Sci. USA 85:861-864.

Morisato, D. and N. Kleckner. 1987. Tn10 transposition and circle formation in vitro. Cell 51:101-111.

Mount, S., M. Green and G. Rubin. 1988. Partial revertants of the transposable element-associated suppressible allele white-apricot in Drosophila melanogaster: structures and responsiveness to genetic modifiers. Genetics 118:221-234.

Nikitin, A. and R. Woodruff. 1995. Somatic movement of the mariner transposable element and lifespan of Drosophila species. Mutat. Res. 338:43-49.

Nuzhdin, S.V. and T. F. C. Mackay. 1995. The genomic rate of transposable element movement in Drosophila melanogaster. Mol. Biol. Evol 12(1):180-181.

Okada, N. 1991. SINEs. Curr. Opin. Genet. Dev. 1:498-504.

Oosumi, T., B. Garlick and W. Belknap. 1995. Identification and characterization of putative transposable DNA elements in solanaceous plants and C. elegans. Proc. Natl. Acad. Sci. USA 92:8886-8890.

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

Orr, H.A. 1995. Somatic Mutation favors the evolution of diploidy. Genetics 139:1441-1447.

Plasterk, R.H.A. 1991. The origin of footprints of the Tc1 transposon of C. elegans. EMBO J 10:1919-1925.

Plasterk, R.H.A. and J.T.M. Groenen. 1992. Targeted alterations of the C. elegans genome by transgene instructed DNA double strand break repair following Tc1 excision. EMBO J. 11:287-290.

Pryciak, P. and H. Varmus. 1992. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. Cell 69:769-780.

Purugganan, M. and S. Wessler. 1992. Splicing of transposable elements and its role in intron evolution. Genetica 86:295-303.

Purugganan, M. 1993. Transposable elements as introns: evolutionary connections. Trends in Ecol. and Evol. 8(7):239-243.

Radice, A. and S. Emmons. 1993. Extrachromosomal circular copies of the transposon Tc1. Nucleic Acids Res. 21(11):2663-2667.

Robertson, H. 1993. The mariner transposable element is widespread in insects. Nature 362:241-245.

Roeder, G.S., and G.R. Fink. 1983. Transposable elements in yeast. Pages 299-326. In J.A. Shapiro (ed.), Mobile Genetic Elements. Academic Press, Inc, New York.

Ronsseray, S. and D. Anxolabehere. 1987. Chromosomal distribution of P and I transposable elements in a natural population of Drosophila melanogaster. Chromosoma 94:433-440.

Rose, A., L Harris, N. Mawji and W. Morris. 1985. Can. J. Biochem. Cell. Biol. 63:752-756.

Rosenzweig, B., L.W. Liao, and D. Hirsh. 1983. Sequence of the C. elegans transposable element Tc1. Nucl. Acids Res. 11:4201-4209.

Ruan, K-S. and S. Emmons. 1984. Extrachromosomal copies of transposon Tc1 in the nematode C. elegans. Proc. Natl. Acad. Sci. USA 81:4018-4022.

Ruan, K-S. and S. Emmons. 1987. Precise and imprecise somatic excision of the transposon Tc1 in the nematode C. elegans. Nucleic Acids Res. 15(17):6875-6881.

Rushforth, A.M., B. Saari and P. Anderson. 1993. Site-selected Insertion of the transposon Tc1 into a C. elegans myosin light chain gene. Mol. and Cell. Biol. 13(2):902-910.

Rushforth, A. and P. Anderson. 1996. Splicing removes the C. elegans Transposon Tc1 from most mutant pre-mRNAs. Mol. and Cell Biol. 16(1):422-429.

Ruvolo, V., J. Hill and A. Levitt. 1992. The Tc2 transposon of C. elegans has the structure of a self-regulated element. DNA Cell Biol. 11:111-122.

Schukkink, R. and R. Plasterk. 1990. TcA, the putative transposase of the C. elegans Tc1 transposon, has an N-terminal DNA binding domain. Nucl. Acids Res. 18(4):895-900.

Schwartz, D. 1984. Analysis of the Ac transposable element dosage effect in maize. Mol. Gen. Genet. 196:81-84.

Schwartz, D. 1989. Pattern of Ac transposition in maize. Genetics 121:125-128.

Shapiro, J. 1992. Natural genetic engineering in evolution. Genetica 86:99-111.

Schneuwly, S., A. Kuroiwa and W. Gehring. 1987. Molecular analysis of the dominant homeotic Antennapedia phenotype. EMBO J. 6:201-206.

Southern, E. 1975. Detection of specific sequences among DNA fragments by gel electrrophoresis. J. Mol. Biol. 98:503-517.

Spradling, A. and G. Rubin. 1982. Transposition of cloned P elements into Drosophila germ-line chromosomes. Science 218:341-347.

Stavenhagen, J. and D. Robins. 1988. An ancient provirus has imposed androgen regulation on the adjacent mouse sex-limited protein. Cell 55:247-254.

Steinmeyer, K., R. Klocke, C. Ortland, M. Gronemeier, H. Jockusch, S. Grunder and T. Jentsch. 1991. Inactivation of muscle chloride channel by transposon insertion in myotonic mice. Nature 354:304-308.

Sulston, J. and H. Horvitz. 1977. Post-embryonic cell lineages of the nematode C. elegans. Dev. Biol. 56:110-156.

Sulston, J., E. Schierenberg, J. White and J. Thompson. 1983. The embryonic cell lineage of the nematode C. elegans. Dev. Biol. 100:64-119.

Swofford, D. 1993. PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1. Computer program distributed by the Illinois Natural History Survey, Champaign, Illinois.

Temin, H. 1980. Origin of retroviruses from cellular movable genetic elements. Cell 21:599-600.

Thierry-Mieg, J. and R. Durbin. 1992. ACeDB, a C. elegans database. cashiers IMBIO 5:15-24.

Torkamanzehi, A., C. Moran and F. Nicholas. 1992. P element transposition contributes substantial new variation for a quantitative trait in Drosophila melanogaster. Genetics 131:73-78.

Tower, J., G. Karpen, N. Craig, A. Spradling. 1993. Preferential insertion of Drosophila P elements to nearby chromosomal sites. Genetics 133:347-359.

van Luenen, H., S. Colloms and R. Plasterk. 1993. Mobilization of quiet, endogenous Tc3 transposons of C. elegans by forced expression of Tc3 transposase. EMBO. J. 12(6):2513-2520.

van Luenen H.G.A.M., R.H.A. Plasterk. 1994. Target site choice of the related transposable elements Tc1 and Tc3 of C. elegans. Nucleic Acids Res. 22(3):262-269.

Vos, J., H. van Luenen and R. Plasterk. 1993. Characterization of the C. elegans Tc1 transposase in vivo and in vitro. Genes Dev. 7:1244-1253.

Walsh, B. 1987. Sequence dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? Genetics 117:543-557.

Wessler, S. 1989. The splicing of maize transposable elements from pre-mRNA- a minireview. Gene 82:127-133.

Wilke, C., E. Maimer and J. Adams. 1993. The population biology and evolutionary significance of Ty elements in S. cerevisiae. Pages 51-69. In Transposable elements and evolution. J. McDonald (ed). Dordrecht: Kluwer Academic Publishers.

Williams, G. 1966. Adaptation and natural selection. Princeton University Press, Princeton, New Jersey.

Wilson, A.C., L.R. Maxson and V.M. Sarich. 1974. Two types of molecular evolution. Evidence from studies of interspecific hybridization. PNAS 71:2843-2847.

Wilson, R., R. Ainscough, K. Anderson, C. Baynes, M. Berks, J. Bonfield, J. Burton, M. Connell, T. Copsey, J. Cooper, A. Coulson, M. Craxton, S. Dear, Z. Du, R. Durbin, A. Favello, L. Fulton, A. Gardner, P. Green, T. Hawkins, L. Hillier, M. Jier, L. Johnston, M. Jones, J. Kershaw, J. Kirsten, N. Laister, P. Latreille, J. Lightning, C. Lloyd, A. McMurray, B. Mortimore, M. O'Callaghan, J. Parsons, C. Percy, L. Rifken, A. Roopra, D. Saunders, R. Shownkeen, N. Smaldon, A. Smith, E. Sonnhammer, R. Staden, J. Sulston, J. Thierry-Mieg, K. Thomas, M. Vaudin, K. Vaughan, R. Waterston, A. Watson, L. Weinstock, J. Wilkinson-Sproat and P. Wohldman. 1994. 2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegans. Nature 368:32-38.

Wood, W. (ed). 1988. The nematode C. elegans. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Woodruff, R. 1993. Transposable DNA elements and life history traits I. Transposition of P DNA elements in somatic cells reduces the lifespan of D. melanogaster. Pages 218-230. In Transposable elements and evolution. J. McDonald (ed.) Kluwer Academic Publishers, Dordrecht, The Netherlands.

Xiong, X. and T. Eickbush. 1990. Origin and Evolution of retroelements based on their reverse transcriptase sequences. EMBO J. 9:3353-3362.

Yuan, J., M. Finney, N. Tsung and R. Horvitz. 1991. Tc4, a C. elegans transposable element with an unusual fold-back structure. Proc. Natl. Acad. Sci. USA 88:3334-3338.

Zwaal, R., A. Broeks, J. Van Meurs, J. Groenen and R. Plasterk. 1993. Target-selected gene inactivation in C. elegans, using a frozen transposon insertion mutant bank. Proc. Natl. Acad. Sci. USA 90:9402-9406.

APPENDIX A: Alignment of Tc1 and seven cosmid sequences identified as high scoring blast hits to Tc1.

```
          1                                                                                                                  100
C28f5rc   GTCTTACCTA CAGTGCTGGC CAAAAAGATA TCCACTTTTG GTTTTTTGTG TGTAACTTTT TTCTCAAGCA TCCATTTGAC TTGAATTTTT CCGTGTGCAT
    Tc1   .......... CAGTGCTGGC CAAAAAGATA TCCACTTTTG GTTTTTTGTG TGTAACTTTT TTCTCAAGCA TCCATTTGAC TTGAATTTTT CCGTGTGCAT
F18c5rc   ACGGGACCTA CAGTGCTGGC CAAAAAGATA TCCACTTTTG GTTTTTTGTG TGTAACTTTT TTCTCAAGCA TCCATTTGAC TTGAATTTTT CCGTGTGCAT
 R173rc   AATATATGTA CAGTGCTGGC CAAAAAGATA TCCACTTTTG GTTTTTTGTG TGTAACTTTT TTCTCAAGCA TCCATTTGAC TTGAATTTTT CCGTGTGCAT
 Zk1251   CATTTGCATA CAGTGCTGGC CAAAAAGATA TCCACTTTTG GTTTTTTGTG TGTAACTTTT TTCTCAAGCA TCCATTTGAC TTGAATTTTT CCGTGTGCAT
 Zk856rc  TCTCCAGGTA CAGTGCTGGC CAAAAAGATA TCCACTTTTG GTTTTTTGTG TGTAACTTTT TTCTCAAGCA TCCATTTGAC TTGAATTTTT CCGTGTGCAT
F08g12rc  GCGGCATGTA CAGTGCTGGC CAAAAAGATA TCCACTTTTG GTTTTTTGTG TGTAACTTTT TTCTCAAGCA TCCATTTGAC TTGAATTTTT CCGTGTGCAT
 R03h10   CACTTATGTA CAGTGCTGGC CAAAAAGATA TCCACTTTTG GTTTTTTGTG TGTAACTTTT TTCTCAAGCA TCCATTTGAC TTGAATTTTT CCGTGTGCAT

          101                                                                                                                200
C28f5rc   AAAGCGAAAT GTTACGCAAA TTTGCGGACC AAACATTACA TGATTATCGA TTTTTTCTGA ATTTTATTTC AATTTTTTGA TTTTTTCGTT TTTCCAATTT
    Tc1   AAAGCGAAAT GTTACGCAAA TTTGCGGACC AAACATTACA TGATTATCGA TTTTTTCTGA ATTTTATTTC AATTTTTTGA TTTTTTCGTT TTTCCAATTT
F18c5rc   AAAGCGAAAT GTTACGCAAA TTTGCGGACC AAACATTACA TGATTATCGA TTTTTTCTGA ATTTTATTTC AATTTTTTGA TTTTTTCGTT TTTCCAATTT
 R173rc   AAAGCGAAAT GTTACGCAAA TTTCCGGACC AAACATTACA TGATTATCGA TTTTTTCTGA ATTTTATTTC AATTTTTTGA TTTTTTCGTT TTTCCAATTT
 Zk1251   AAAGCGAAAT GTTACGCAAA TTTGCGGACC AAACATTACA TGATTATCGA TTTTTTCTGA ATTTTATTTC AATTTTTTGA TTTTTTCGTT TTTCCAATTT
 Zk856rc  AAAGCGAAAT GTTACGCAAA TTTGCGGACC AAACATTACA TGATTATCGA TTTTTTCTGA ATTTTATTTC AATTTTTTGA TTTTTTCGTT TTTCCAA.TT
F08g12rc  AAAGCGAAAT GTTACGCAAA TTTGCGGACC AAACATTACA TGATTATCGA TTTTTTCTGA ATTTTATTTC AATTTTTTGA TTTTTTCGTT TTTCCAATTT
 R03h10   AAAGCGAAAT GTTACGCAAA TTTGCGGACC GAACATTACA TGATTATCGA TTTTTTCTGA ATTTTATTTC AATTTTTTGA TTTTTTCGTT TTTCCAATTT

          201                                                                                                                300
C28f5rc   TCATTATTTT TTTTGAATTA TCAATAAAAC GCACTCTGTT TGTTGCACTG GATTTGTTTG GTTGATAAAT TATTTTTAAG GTATGGTAAA ATCTGTTGGG
    Tc1   TCATTATTTT TTTTGAATTA TCAATAAAAC GCACTCTGTT TGTTGCACTG GATTTGTTTG GTTGATAAAT TATTTTTAAG GTATGGTAAA ATCTGTTGGG
F18c5rc   TCATTATTTT TTTTGAATTA TCAATAAAAC GCACTCTGTT TGTTGCACTG GATTTGTTTG GTTGATAAAT TATTTTTAAG GTATGGTAAA ATCTGTTGGG
 R173rc   TCATTATTTT TTTTGAATTA TCAATAAAAC GCACTCTGTT TGTTGCACTG GATTTGTTTG GTTGATAAAT TATTTTTAAG GTATGGTAAA ATCTGTTGGG
 Zk1251   TCATTATTTT TTTTGAATTA TCAATAAAAC GCACTCTGTT TGTTGCACTG GATTTGTTTG GTTGATAAAT TATTTTTAAG GTATGGTAAA ATCTGTTGGG
 Zk856rc  TCATTATTTT TTTTGAATTA TCAATAAAAC GCACTCTGTT TGTTGCACTG GATTTGTTTG GTTGATAAAT TATTTTTAAG GTATGGTAAA ATCTGTTGGG
F08g12rc  TCATTATTTT TTTTGAATTA TCAATAAAAC GCACTCTGTT TGTTGCACTG GATTTGTTTG GTTGATAAAT TATTTTTAAG GTATGGTAAA ATCTGTTGGG
 R03h10   TCATTATTTT TTTTGAATTA TCAATAAAAC GCACTCTGTT TGTTGCACTG GATTTGTTTG GTTGATAAAT TATTTTTAAG GTATGGTAAA ATCTGTTGGG
```

```
          301                                                                                                   400
C28f5rc   TGTAAAAATC TTTCCTTGGA CGTCAAGAAA GCCATTGTAG CTGGCTTCGA ACAAGGAATA CCCACGAAAA TGCTCGCGCT GCAAATTCAA CGTTCTCCGT
    Tc1   TGTAAAAATC TTTCCTTGGA CGTCAAGAAA GCCATTGTAG CTGGCTTCGA ACAAGGAATA CCCACGAAAA .GCTCGCGCT GCAAATTCAA CGTTCTCCGT
F18c5rc   TGTAAAAATC TTTCCTTGGA CGTCAAGAAA GCCATTGTAG CTGGCTTCGA ACAAGGAATA CCCACGAAAA TGCTCGCGCT GCAAATTCAA CGTTCTCCGT
 R173rc   TGTAAAAATC TTTCCTTGGA CGTCAAGAAA GCCATTGTAG CTGGCTTCGA ACAAGGAATA CCCACGAAAA TGCTCGCGCT GCAAATTCAA CGTTCTCCGT
 Zk1251   TGTAAAAATC TTTCCTTGGA CGTCAAGAAA GCCATTGTAG CTGGCTTCGA ACAAGGAATA CCCACGAAAA TGCTCGCGCT GCAAATTCAA CGTTCTCCGT
Zk856rc   TGTAAAAATC TTTCCTTGGA CGTCAAGAAA GCCATTGTAG CTGGCTTCGA ACAAGGAATA CCCACGAAAA TGCTCGCGCT GCAAATTCAA CGTTCTCCGT
F08g12rc  TGTAAAAATC TTTCCTTGGA CGTCAAGAAA GCCATTGTAG CTGGCTTCGA ACAAGGAACA CCCACGAAAA TGCTCGCGCT GCAAATTCAA CGTTCTCCGT
 R03h10   TGTAAAAATC TTTCCTTGGA CGTCAAGAAA GCCATTGTAG CTGGCTTCGA ACAAGGAATA CCCACGAAAA TGCTCGCGCT GCAAATTCAA CGTTCTCCGT

          401                                                                                                   500
C28f5rc   CGACTATTTG GAAAGTAATC AAGAAGTACC AAACTGAGGT GAGTTCGAAA AATATTATTT TTTAATAATA AATGTTTAGA AATCCGTCGC TTTGAGAATC
    Tc1   CGACTATTTG GAAAGTAATC AAGAAGTACC AAACTGAGGT GAGTTCGAAA AATATTATTT TTTAATAATA AATGTTTAGA AATCCGTCGC TTTGAGAATC
F18c5rc   CGACTATTTG GAAAGTAATC AAGAAGTACC AAACTGAGGT GAGTTCGAAA AATATTATTT TTTAATAATA AATGTTTAGA AATCCGTCGC TTTGAGAATC
 R173rc   CGACTATTTG GAAAGTAATC AAGAAGTACC AAACTGAGGT GAGTTCGAAA AATATTATTT TTTAATAATA AATGTTTAGA AATCCGTCGC TTTGAGAATC
 Zk1251   CGACTATTTG GAAAGTAATC AAGAAGTACC AAACTGAGGT GAGTTCGAAA AATATTATTT TTTAATAATA AATGTTTAGA AATCCGTCGC TTTGAGAATC
Zk856rc   CGACTATTTG GAAAGTAATC AAGAAGTACC AAACTGAGGT GAGTTCGAAA AATATTATTT TTTAATAATA AATGTTTAGA AATCCGTCGC TTTGAGAATC
F08g12rc  CGACTATTTG GAAAGTAATC AAGAAGTACC AAACTGAGGT GAGTTCGAAA AATATTATTC TTTAATAATA AATGTTTAGA AATCCGTCGC TTTGAGAATC
 R03h10   CGACTATTTG GAAAGTAATC AAGAAGTACC AAACTGAGGT GAGTTCGAAA AATATTATTT TTTAATAATA AATGTTTAGA AATCCGTCGC TTTGAGAATC

          501                                                                                                   600
C28f5rc   TCGCCCGGCA GGCCTCGAGT GACAACCCAT AGGATGGATC GCAACATCCT CCGATCAGCA AGAGAAGATC CGCATAGGAC CGCCACGGAT ATTCAAATGA
    Tc1   TCGCCCGGCA GGCCTCGAGT GACAACCCAT AGGATGGATC GCAACATCCT CCGATCAGCA AGAGAAGATC CGCATAGGAC CGCCACGGAT ATTCAAATGA
F18c5rc   TCGCCCGGCA GGCCTCGAGT GACAACCCAT AGGATGGATC GCAACATCCT CCGATCAGCA AGAGAAGATC CGCATAGGAC CGCCACGGAT ATTCAAATGA
 R173rc   TCGCCCGGCA GGCCTCGAGT GACAACCCAT AGGATGGATC GCAACATCCT CCGATCAGCA AGAGAAGATC CGCATAGGAC CGCCACGGAT ATTCAAATGA
 Zk1251   TCGCCCGGCA GGCCTCGAGT GACAACCCAT AGGATGGATC GCAACATCCT CCGATCAGCA AGAGAAGATC CGCATAGGAC CGCCACGGAT ATTCAAATGA
Zk856rc   TCGCCCGGCA GGCCTCGAGT GACAACCCAT AGGATGGATC GCAACATCCT CCGATCAGCA AGAGAAGATC CGCATAGGAC CGCCACGGAT ATTCAAATGA
F08g12rc  TCGCCCGGCA GGCCTCGAGT GACAACCCAT AGGATGGATC GCAACATCCT CCGATCAGCA AGAGAAGATC CGCATAGGAC CGCCACGGAT ATTCAAACGA
 R03h10   TCGCCCGGCA GGCCTCGAGT GACAACCCAT AGGATGGATC GCAACATCCT CCGATCAGCA AGAGAAGATC CGCATAGGAC CGCCACGGAT ATTCAAATGA
```

```
           601                                                                                                     700
 C28f5rc   TTATAAGTTC TCCAAATGAA CCTGTACCAA GTAAACGAAC TGTTCGTCGA CGTTTACAGC AAGCAGGACT ACATGGACGA AAGCCAGTCA AGAAACCGTT
     Tc1   TTATAAGTTC TCCAAATGAA CCTGTACCAA GTAAACGAAC TGTTCGTCGA CGTTTACAGC AAGCAGGACT ACACGGACGA AAGCCAGTCA AGAAACCGTT
 F18c5rc   TTATAAGTTC TCCAAATGAA CCTGTACCAA GTAAACGAAC TGTTCGTCGA CGTTTACAGC AAGCAGGACT ACACGGACGA AAGCCAGTCA AGAAACCGTT
  R173rc   TTATAAGTTC TCCAAATGAA CCTGTACCAA GTAAACGAAC TGTTCGTCGA CGTTTACAGC AAGCAGGACT ACACGGACGA AAGCCAGTCA AGAAACCGTT
  Zk1251   TTATAAGTTC TCCAAATGAA CCTGTACCAA GTAAACGAAC TGTTCGTCGA CGTTTACAGC AAGCAGGACT ACACGGACGA AAGCCAGTCA AGAAACCGTT
 Zk856rc   TTATAAGTTC TCCAAATGAA CCTGTACCAA GTAAACGAAC TGTTCGTCGA CGTTTACAGC AAGCAGGACT ACACGGACGA AAGCCAGTCA AGAAACCGTT
F08g12rc   TTATAAGTTC TCCAAATGAA CCTGTACCAA GTAAACGAAC TGTTCGTCGA CGTTTACAGC AAGCAGGACT ACACGGACGA AAGCCAGTCA AGAAACCGTT
  R03h10   TTATAAGTTC TCCAAATGAA CCTGTACCAA GTAAACGAAC TGTTCGTCGA CGTTTACAGC AAGCAGGACT ACACGGACGA AAGCCAGTCA AGAAACCGTT

           701                                                                                                     800
 C28f5rc   CATCAGTAAG AAAAATCGCA TGGCTCGAGT TGCGTGGGCA AAAGCGCATC TTCGTTGGGG ACGTCAGGAA TGGGCTAAAC ACATCTGGTC TGACGAAAGC
     Tc1   CATCAGTAAG AAAAATCGCA TGGCTCGAGT TGCGTGGGCA AAAGCGCATC TTCGTTGGGG ACGTCAGGAA TGGGCTAAAC ACATCTGGTC TGACGAAAGC
 F18c5rc   CATCAGTAAG AAAAATCGCA TGGCTCGAGT TGCGTGGGCA AAAGCGCATC TTCGTTGGGG ACGTCAGGAA TGGGCTAAAC ACATCTGGTC TGACGAAAGC
  R173rc   CATCAGTAAG AAAAATCGCA TGGCTCGAGT TGCGTGGGCA AAAGCGCATC TTCGTTGGGG ACGTCAGGAA TGGGCTAAAC ACATCTGGTC TGACGAAAGC
  Zk1251   CATCAGTAAG AAAAATCGCA TGGCTCGAGT TGCGTGGGCA AAAGCGCATC TTCGTTGGGG ACGTCAGGAA TGGGCTAAAC ACATCTGGTC TGACGAAAGC
 Zk856rc   CATCAGTAAG AAAAATCGCA TGGCTCGAGT TGCGTGGGCA AAAGCACATC TTCGTTGGGG ACGTCAGGAA TGGGCTAAAC ACATCTGGTC TGACGAAAGC
F08g12rc   CATCAGTAAG AAAAATCGCA TGGCTCGAGT TGCGTGGGCA AAAGCGCATC TTCGTTGGGG ACGTCAGGAA TGGGCTAAAC ACATCTGGTC TGACGAAAGC
  R03h10   CATCAGTAAG AAAAATCGCA TGGCTCGAGT TGCGTGGGCA AAAGCGCATC TTCGTTGGGG ACGTCAGGAA TGGGCTAAAC ACATCTGGTC TGACGAAAGC

           801                                                                                                     900
 C28f5rc   AAGTTCAATT TGTTCGGGAG TGATGGAAAT TCCTGGGTAC GTCGTCCTGT TGGCTCTAGG TACTCTCCAA AGTATCAATG CCCAACCGTT AAGCATGGAG
     Tc1   AAGTTCAATT TGTTCGGGAG TGATGGAAAT TCCTGGGTAC GTCGTCCTGT TGGCTCTAGG TACTCTCCAA AGTATCAATG CCCAACCGTT AAGCATGGAG
 F18c5rc   AAGTTCAATT TGTTCGGGGG TGATGGAAAT TCCTGGGTAC GTCGTCCTGT TGGCTCTAGG TACTCTCCAA AGTATCAATG CCCAACCGTT AAGCATGGAG
  R173rc   AAGTTCAATT TGTTCGGGAG TGATGGAAAT TCCTGGTTAC GTCGTCCTGT TGGCTCTAGG TACTCTCCAA AGTATCAATG CCCAACCGTT AAGCATGGAG
  Zk1251   AAGTTCAATT TGTTCGGGAG TGATGGAAAT TCCTGGGTAC GTCGTCCTGT TGGCTCTAGG TACTCTCCAA AGTATCAATG CCCAACCGTT AAGCATGGAG
 Zk856rc   AAGTTCAATT TGTTCGGGAG TGATGGAAAT TCCTGGGTAC GTCGTCCTGT TGGCTCTAGG TACTCTCCAA AGTATCAATG CCCAACCGTT AAGCATGGAG
F08g12rc   AAGTTCAATT TGTTCGGGAG TGATGGAAAT TCCTGGGTAC GTCGTCCTGT TGGCTCTAGG TACTCTCCAA AGTATCAATG CCCAACCGTT AAGCATGGAG
  R03h10   AAGTTCAATT TGTTCGGGAG TGATGGAAAT TCCTGGGTAC GTCGTCCTGT TGGCTCTAGG TACTC..... .......... .......... ..........
```

```
          901                                                                                                                  1000
C28f5rc   GTGGGAGCGT CATGGTGTGG GGGTGCTTCA CCAGCACTTC CATGGGCCCA CTAAGGAGAA TCCAAAGCAT TATGGATCGT TTTCAATACG AAAACATCTT
    Tc1   GTGGGAGCGT CATGGTGTGG GGGTGCTTCA CCAGCACTTC CATGGGCCCA CTAAGGAGAA TCCAAAGCAT TATGGATCGT TTTCAATACG AAAACATCTT
F18c5rc   GTGGGAGCGT CATGGTGTGG GGGTGCTTCA CCAGCACTTC CATGGGCCCA CTAAGGAGAA TCCAAAGCAT TATGGATCGT TTTCAATACG AAAACATCTT
 R173rc   GTGGGAGCGT CATGGTGTGG GGGTGCTTCA CCAGCACTTC CATGGGCCCA CTAAGGAGAA TCCAAAGCAT TATGGATCGT TTTCAATACG AAAACATCTT
 Zk1251   GTGGGAGCGT CATGGTGTGG GGGTGCTTCA CCAGCACTTC CATGGGCCCA CTAAGGAGAA TCCAAAGCAT TATGGATCGT TTTCAATACG AAAACATCTT
Zk856rc   GTGGGAGCGT CATGGTGTGG GGGTGCTTCA CCAGCACTTC CATGGGCCCA CTAAGGAGAA TCCAAAGCAT TATGGATCGT TTTCAATACG AAAACATCTT
F08g12rc  GTGGGAGCGT CATGGTGTGG GGGTGCTTCA CCAGCACTTC CATGGGCCCA CTAAGGAGAA TCCAAAGCAT TATGGATCGT TTTCAATACG AAAACATCTT
 R03h10   .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........

          1001                                                                                                                 1100
C28f5rc   GGAAACTACA ATGCGACCCT GGGCACTTCA AAATGTGGGC CGTGGCTTCG TGTTTCAGCA GGATAACGAT CCTAAGCATA CTTCTCTTCA TGTGCGTTCC
    Tc1   TGAAACTACA ATGCGACCCT GGGCACTTCA AAATGTGGGC CGTGGCTTCG TGTTTCAGCA GGATAACGAT CCTAAGCATA CTTCTCTTCA TGTGCGTTCA
F18c5rc   GGAAACTACA ATGCGACCCT GGGCACTTCA AAATGTGGGC CGTGGCTTCG TGTTTCAGCA GGATAACGAT CCTAAGCATA CTTCTCTTCA TGTGCGTTCC
 R173rc   GGAAACTACA ATGCGACCCT GGGCACTTCA AAATGTGGGC CGTGGCTTCG TGTTTCAGCA GGATAACGAT CCTAAGCATA CTTCTCTTCA TGTGCGTTCC
 Zk1251   GGAAACTACA ATGCGACCCT GGGCACTTCA AAATGTGGGC CGTGGCTTCG TGTTTCAGCA GGATAACGAT CCTAAGCATA CTTCTCTTCA TGTGCGTTCC
Zk856rc   GGAAACTACA ATGCGACCCT GGGCACTTCA AAATGTGGGC CGTGGCTTCG TGTTTCAGCA GGATAACGAT CCTAAGCATA CTTCTCTTCA TGTGCGTTCC
F08g12rc  GGAAACTACA ATGCGACCCT GGGCACTTCA AAATGTGGGC CGTGGCTTCG TGTTTCAGCA GGATAACGAT CCTAAGCATA CTTCTCTTCA TGTGCGTTCC
 R03h10   .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........

          1101                                                                                                                 1200
C28f5rc   TGGTTTCAAC GTCGTCGTGT GCATTTGCTC GATTGGCCAA GTCAGTCTCC GGACTTGAAT CCAATAGAGC ATTTGTGGGA AGAGTTGGAA AGACGTCTTG
    Tc1   TGGTTTCAAC GTCGTCATGT GCATTTGCTC GATTGGCCAA GTCAGTCTCC GGACTTGAAT CCAATAGAGC ATTTGTGGGA AGAGTTGGAA AGACGTCTTG
F18c5rc   TGGTTTCAAC GTCGTCGTGT GCATTTGCTC GATTGGCCAA GTCAGTCTCC GGACTTGAAT CCAATAGAGC ATTTGTGGGA AGAGTTGGAA AGACGTCTTG
 R173rc   TGGTTTCAAC GTCGTCGTGT GCATTTGCTC GATTGGCCAA GTCAGTCTCC GGACTTGAAT CCAATAGAGC ATTTGTGGGA AGAGTTGGAA AGACGTCTTG
 Zk1251   TGGTTTCAAC GTCGTCGTGT GCGTTTGCTC GATTGGCCAA GTCAGTCTCC GGACTTGAAT CCAATAGAGC ATTTGTGGGA AGAGTTGGAA AGACGTCTTG
Zk856rc   TGGTTTCAAC GTCGTCGTGT GCATTTGCTC GATTGGCCAA GTCAGTCTCC GGACTTGAAT CCAATAGAGC ATTTGTGGGA AGAGTTGGAA AGACGTCTTG
F08g12rc  TGGTTTCAAC CTTGTCGTGT GCATTTGCTC GATTGGCCAA GTCAGTCTCC GGACTTGAAT CCAATAGAGC ATTTGTGGGA AGAGTTGGAA AGACGTCTTG
 R03h10   .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
```

```
          1201                                                                                                                  1300
C28f5rc   GAGGTATTCG GGCTTCAAAT GCAGATGCCA AATTCAACCA GTTGGAAAAC GCTTGGAAAG CTATCCCCAT GTCAGTTATT CACAAGCTGA TCGACTCGAT
    Tc1   GAGGTATTCG GGCTTCAAAT GCAGATGCCA AATTCAACCA GTTGGAAAAC GCTTGGAAAG CTATCCCCAT GTCAGTTATT CACAAGCTGA TCGACTCGAT
F18c5rc   GAGGTATTCG GGCTTCAAAT GCAGATGCCA AATTCAACCA GTTGGAAAAC GCTTGGAAAG CTATCCCCAT GTCAGTTATT CACAAGCTGA TCGACTCGAT
 R173rc   GAGGTATTCG GGCTTCAAAT GCAGATGCCA AATTCAACCA GTTGGAAAAC GCTTGGAAAG CTATCCCCAT GTCAGTTATT CACAAGCTGA TCGACTCGAT
 Zk1251   GAGGTGTTCG GGCTTCAAAT GCAGATGCCA AATTCAACCA GTTGGAAAAC GCTTGGAAAG CTATCCCCAT GTCAGTTATT CACAAGCTGA TCGACTCGAT
Zk856rc   GAGGTATTCG GGCTTCAAAT GCAGATGCCA AATTCAACCA GTTGGAAAAC GCTTGGAAAG CTATCCCCAT GTCAGTTATT CACAAGCTGA TCGACTCGAT
F08g12rc  GAGGTATTCG GGCTTCAAAT GCAGATGCCA AATTCAACCA GTTGGAAAAC GCTTGGAAAG CTATCCCCAT GTCAGTTATT CACAAGCTGA TCGACTCGAT
 R03h10   .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........

          1301                                                                                                                  1400
C28f5rc   GCCACGTCGT TGTCAAGCTG TTATTGATGC AAACGGATAC GCGACAAAGT ATTAAGCATA ATTATGTTGT TTTTAAATCC AATTGCTCAT ATTCCGGTAC
    Tc1   GCCACGTCGT TGTCAAGCTG TTATTGATGC AAACGGATAC GCGACAAAGT ATTAAGCATA ATTATGTTGT TTTTAAATCC AATTGCTCAT ATTCCGGTAC
F18c5rc   GCCACGTCGT TGTCAAGCTG TTATTGATGC AAACGGATAC GCGACAAAGT ATTAAGCATA ATTATGTTGT TTTTAAATCC AATTGCTCAT ATTCCGGTAC
 R173rc   GCCACGTCGT TGTCAAGCTT TTATTGATGC AAACGGATAC GCGACAAAGT ATTAAGCATA ATTATGTTGT TTTTAAATCC AATTGCTCAT ATTCCGGTAC
 Zk1251   GCCACGTCGT TGTCAAGCTG TTATTGATGC AAACGGATAC GCGACAAAGT ATTAAGCATA ATTATGTTAT TTTTAAATCC AATTGCTCAT ATTCCGGTAC
Zk856rc   GCCACGTCGT TGTCAAGCTG TTATTGATGC AAACGGATAC GCGACAAAGT ATTAAGCATA ATTATGTTGT TTTTAAATCC AATTGCTCAT ATTCCGGTAC
F08g12rc  GCCACGTCGT TGTCTAGCTG TTATTGATGC AAACGGATAC GCGACAAAGT ATTAAGCATA ATTATGTTGT TTTTAAATCC AATTGCTCAT ATTCCGGTAC
 R03h10   .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........

          1401                                                                                                                  1500
C28f5rc   TTTAATTGTC ATTTCCTTGC AATCTCGGTT TTTTCAATAT TTCTAGTTTT TCGATTTTTT TGAATTTTTC TGAAGTTTTT TCAAAATCTG TTGAACATTT
    Tc1   TTTAATTGTC ATTTCCTTGC AACCTCGGTT TTTTCAATAT TTCTAGTTTT TCGATTTTTT TGAATTTTTC TGAAGTTTTT TCAAAATCTG TTGAACATTT
F18c5rc   TTTAATTGTC ATTTCCTTGC AATCTCGGTT TTTTCAATAT TTCTAGTTTT TCGATTTTTT TGAATTTTTC TGAAGTTTTT TCAAAATCTG TTGAACATTT
 R173rc   TTTAATTGTC ATTTCCTTGC AACCTCGGTT TTTTCAATAT TTCTAGTTTT TCGATTTTTT TGAATTTTTC TGAAGTTTTT TCAAAATCTG TTGAACATTT
 Zk1251   TTTAATTGTC ATTTCCTTGC AACCTCGGTT TTTTCAATAT TTCTAGTTTT TCGATTTTTT TGAATTTTTC TGAAGTTTTT TCAAAATCTG TTGAACATTT
Zk856rc   TTTAATTGTC ATTTCCTTGC AACCTCGGTT TTTTCAATAT TTCTAGTTTT TCGATTTTTT TGAATTTTTC TGAAGTTTTT TCAAAATCTG TTGAACATTT
F08g12rc  TTTAATTGTC ATTTCCTTGC AACCTCGGTT TTTTCAATAT TTCTAGTTTT TCGATTTTTT TGAATTTTTC TGAAGTTTTT TCAAAATCTG TTGAACATTT
 R03h10   .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
```

```
         1501                                                                                                                1600
C28f5rc  TTGATGAATA TTGTGTTTTT AGAATTTGTG AACACTGTGG TGAAGTTTCA AAACAAAATA ACCACTTAGA AAAAAGTTAC ACACAAAAAA CCAAAAGTGG
    Tc1  TTGATGAATA TTGTGTTTTT AGATTTTGTG AACACTGTGG TGAAGTTTCA AAACAAAATA ACCACTTAGA AAAAAGTTAC ACACAAAAAA CCAAAAGTGG
F18c5rc  TTGATGAATA TTGTGTTTTT AGAATTTGTG AACACTGTGG TGAAGTTTCA AAACAAAATA ACCACTTAGA AAAAAGTTAC ACACAAAAAA CCAAAAGTGG
 R173rc  TTGATGAATA TTGTGTTTTT AGATTTTGTG AACACTGTGG TGAAGTTTCA AAACAAAATA ACCACTTAGA AAAAAGTTAC ACACAAAAAA CCAAAAGTGG
 Zk1251  TTGATGAATA TTGTGTTTTT AGATTTTGTG AACACTGTGG TGAAGTTTCA AAACAAAATA ACCACTTAGA AAAAAGTTAC ACACAAAAAA CCAAAAGTGG
Zk856rc  TTGATGAATA TTGTGTTTTT AGATTTTGTA AACACTGTGG TGAAGTTTCA AAACAAAATA ACCACTTAGA AAAAAGTTAC ACACAAAAAA CCAAAAGTGG
F08g12rc TTGATGAATA TTGTGTTTTT AGATTTTGTG AACACTGTGG TGAAGTTTCA AAACAAAATA ACCACTTAGA AAAAAGTTAC ACACAAAAAA CCAAAAGTGG
 R03h10  .......... .......... .......... .......... .......TCA AAACAAAATA ACCACTTAGA AAAAAGTTAC ACACAAAAAA CCAAAAGTGG

         1601                        1631
C28f5rc  ATATCTTTTT GGCCAGCACT GTATATGTTG T
    Tc1  ATATCTTTTT GGCCAGCACT G......... .
F18c5rc  ATATCTTTTT GGCCAGCACT GTATCTCTTG A
 R173rc  ATATCTTTTT GGCCAGCACT GTATGAATCA T
 Zk1251  ATATCTTTTT GGCCAGCACT GTACCTGCAC A
Zk856rc  ATATCTTTTT GGCCAGCACT GTAAGTGTTA C
F08g12rc ATATCTTTTT GGCCAGCACT GTAACACACA T
 R03h10  ATATCTTTTT GGCCAGCACT GTAAATTTCA A
```

APPENDIX B: Alignment of seven additional cosmid sequences identified as high scoring blast hits to Tc1.

```
         1                                                                                                                  100
C07d10   CTAATATCTA CAGTGCTGGC CAAAAAGATA TCCACTTTCA GTTTTTTGAC GATTTCGATA TTTTTTCCAA TGGGCATAAC TTCAAAACTA GGAAAGGTAC
Zk899    TACCGACATA CAGTGCTGGC CAAAAAGATA TCCACTTTCA GTTTTTTGAC GATTTCGATA TTTTTTCCAA TGGGCATAAC TTCAAAACTA GGAAAGGTAC
D1009rc  TTCGCACATA CAGTGCTGGC CAAAAAGATA TCCACTTTCA GTTTTTTGAC GATTTCGATA TTTTTTCCAA TGGGCATAAC TTCAAAACTA GGAAAGGTAC
M02d8    TATATATATA CAGTGCTGGC CAAAAAGATA TCCACTTTCA GTTTTTTGAC GATTTCGATA TTTTTTCCAA TGGGCATAAC TTCAAAACTA GGAAAGGTAC
F02d10   GAATAACATA CAGTGCTGGC CAAAAAGATA TCCACTTTCA GTTTTTTGAC GATTTCGATA TTTTTTCCAA TGGGCATAAC TTCAAAACTA GGAAAAGTAC
F19h6    CAAACATATA CAGTGCTGGC CAAAAAGATA TCCACTTTCA GTTTTTTGAC GATTTCGATA TTTTTTCCAA TGGGCATAAC TTCAAAACTA GGAAAGGTAC
C30g4    TCACAAAGTA CAGTGCTGGC CAAAATGATA TCCACTCTTA GTTTTTTGAT GATTTCGTTA TTTTTTCCGA TGAGTGTAAC TTAAAAACTA AAAATGCTAT

         101                                                                                                                200
C07d10   CAAAAAATTT TCAACTGGTA AAATGTAGCT CGTGATCAGG CCTATGTATT TTTACATGTT GCAATTATTC ATCCATCACA TGGCAAGTAA TAAAGCGGCG
Zk899    CAAAAAATTT TCAACTGGTA AAATGTAGCT CGTGATCAGG CCTATGTATT TTTACATGTT GCAATTATTC ATCCATCACA TGGCAAGTAA TAAAGCGGCG
D1009rc  CAAAAAATTT TCAACTGGTA AAATGTAGCT CGTGATCAGG CCTATGTATT TTTACATGTT GCAATTATTC ATCCATCACA TGGCAAGTAA TAAAGCGGCG
M02d8    CAAAAAATTT TCAACTGGTA AAATGTAGCT CGTGATCAGG CCTATGTATT TTTACATGTT GCAATTATTC ATCCATCACA TGGCAAGTAA TAAAGCGGCG
F02d10   CAAAAAATTT TCAACTGGTA AAATGTAGCT CGTGATCAGG CCTATGTATT TTTACATGTT GCAATTATTC ATCCATCACA TGGCAAGTAA TAAAGCGGCG
F19h6    CAAAAAATTT TCAACTGGTA AAATGTAGCT CGTGATCAGG CCTATGTATT TTTACATGTT GCAATTAT.. .......... .......... ..........
C30g4    CAACAAATTT TCAACTGGGA AAGTATAGCC CGTGATCAGG TGTATTTCTT TTTACATGTT TGAAAAAATC AATAAAATCA TGGCAAGAAA TAAAGCGGCC

         201                                                                                                                300
C07d10   GGCATCTCGT GAGTCCGTTT TTGACGATGA TTACTAAAAC GACTGTAACT CAAGAAACAT ATTTTTAATG AAAGGTTTGA GAAAGTAACA AAATGTTTAT
Zk899    GGCATCTCGT GAGTCCGTTT TTGACGATGA TTACTAAAAC GACTGTAACT CAAGAAACAT ATTTTTAATG AAAGGTTTGA GAAAGTAACA AAATGTTTAT
D1009rc  GGCATCTCGT GAGTCCGTTT TTGACGATGA TTACTAAAAC GACTGTAACT CAAGAAACAT ATTTTTAATG AAAGGTTTGA GAAAGTAACA AAATGTTTAT
M02d8    GGCATCTCGT GAGTCCGTTT TTGACGATGA TTACTAAAAC GACTGTAACT CAAGAAACAT ATTTTTAATG AAAGGTTTGA GAAAGTAACA AAATGTTTAT
F02d10   GGCATCTCGT GAGTCCGTTT TTGACGATGA TTACTAAGAC GACTGTAACT CAAGAAACAT ATTTTTAATG AAAGGTTTGA GAAAGTAACA AAATGTTTAT
F19h6    .......... ..GTCCGTTT TTGACGATGA TTACTAAAAC GACTGTAACT CAAGAAACAT ATTTTTAATG AAAGGTTTGA GAAAGTAACA AAATGTTTAT
C30g4    GACATCTCGT GAGTCCATTT TTGATGATGG TTACGAAAAC GACTGTAACT CAAGAGCTAT ATTTTTAATG GAAGGTTTGT GAAAG..... ..........

         301                                                                                                                400
C07d10   TTAATTTTTC ATTGTTTGAA CATATCAACT TTGTCCTAAA ACCTCCATTT AAAAAAATGT ATGCGCTGAA ACTAGTGTCT CATTAGACAC TGTTTAGAGG
Zk899    TTAATTTTTC ATTGTTTGAA CATATCAACT TTGTCCTAAA ACCTCCATTT AAAAAAATGT ATGCGCTGAA ACTAGTGTCT CATTAGACAC TGTTTAGAGG
D1009rc  TTAATTTTTC ATTGTTTGAA CATATCAACT TTGTCCTAAA ACCTCCATTT AAAAAAATGT GTGCGCTGAA ACTAGTGTCT CATTAGACAC TGTTTAGAGG
M02d8    TTAATTTTTC ATTGTTTGAA CATATCAACT TTGTCCTAAA ACCTCCATTT AAAAAAATGT GTGCGCTGAA ACTAGTGTCT CATTAGACAC TGTTTAGAGG
F02d10   TTAATTTTTC ATTGTTTGAA CATATCAACT TTGTCCTAAA ACCTCCATTT AAAAAAATGT GTGCGCTGAA ACTAGTGTCT CATTAGACAC TGTTTAGAGG
F19h6    TTAATTTTTT ATTGTTTGAA CATATCAACT TTGTCCTAAA ACCTCCATTT AAAAAAATGT GTGCGCTGAA ACTAGTGTCT CATTAGACAC TGTTTAGAGG
C30g4    .......... .......... .......... .......... ...TCTATCA TAAACAGTGG AT........ .......... .......... ..........
```

```
          401                                                                                                           500
C07d10    CTTTGTTCAA AAATCAGGTT TCTTGGATTG AAAATCTTTT TCCGACAATT TTTGTGAAGA ACTGATGTTG TTATTATATT TGAACTACAT ATTAACCAAT
 Zk899    CTTTGTTCAA AAATCAGGTT TCTTGGATTG AAAATCTTTT TCCGACAATT TTTGTGAAGA ACTGATGTTG TTATTATATT TGAACTACAT ATTAACCAAT
D1009rc   CTTTGTTCAA AAATCAGGTT TCTTGGATTG AAAATCTTTT TCCGACAATT TTTGTGAAGA ACTGATGTTG TTATTATATT TGAACTACAT ATTAACCAAT
 M02d8    CTTTGTTCAA AAATCAGGTT TCTTGGATTG AAAATCTTTT TCCGACAATT TTTGTGAAGA ACTGATGTTG TTATTATATT TGAACTACAT ATTAACCAAT
F02d10    CTTTGTTCAA AAATCAGGTT TCTTGGATTG AAAATCTTTT TCCGACAATT TTTGTGAAGA ACTGATGTTG TTATTATATT TGAACTACAT ATTAACCAAT
 F19h6    CTTTGTTCAA AAATCAGGTT TCTTGGATTG AAAATCTTTT TCCGACAATT TTTGTGAAGA ACTGATGTTG TTATTATATT TGAACTACAT ATTAACCAAT
 C30g4    AATTTTTTCT GAATTTTATT TTTTGGTTAG TATAT..... .......... .......AGT TCAAATATCA TAACTACATT AGTTCTTCAC AAAACTGATC

          501                                                                                                           600
C07d10    AAACAGAATT CAGAA.AAAA .AATCTACTA TTTTTGAGAC ATGAGCTCAA GATTACTCTA CACTTCTCTT CTTTAAACTC TCTTGTCTTT TTCAATGGAG
 Zk899    AAACAGAATT CAGAA.AAAA .AATCTACTA TTTTTGAGAC ATGAGCTCAA GATTACTCTA CACTTCTCTT CTTTAAACTC TCTTGTCTTT TTCAATGGAG
D1009rc   AAACAGAATT CAGAA.AAAA AAATCTACTA TTTTTGAGAC ATGAGCTCGA GATTACTCTA CACTTCTCTT CTTTAAACTC TCTTGTCTTT TTCAATGGAG
 M02d8    AAACAGAATT CAGAA.AAAA AAATCTACTA TTTTTGAGAC ATGAGCTCGA GATTACTCTA CACTTCTCTT CTTTAAACTC TCTTGTCTTT TTCAATGGAG
F02d10    AAACAGAATT CAGAA.AAAA AAATCTACTA TTTTTGAGAC ATGAGCTCGA GATTACTCTA CACTTCTCTT CTTTAAACTC TCTTGTCTTT TTCAATGGAG
 F19h6    AAACAGAATT CAGAA.AAAA AAATCTACTA TTTTTGAGAC ATGAGCTCGA GATTACTCTA CACTTCTCTT CTTTAAACTC TCTTGTCTTT TTCAATGGAG
 C30g4    AAAAAAAATT TTCAATCCAA GAATCCTGAT TTTTGAATAT GCCGTCTAAA CAGTGTCTTG TGAGGCACTA GTCTCAGCTC GCATATTTTT TTAAATCGAG

          601                                                                                                           700
C07d10    GTTTCAGGAC GAAGTTAGTA TGTTCAAACA AT..AAAAAA TTATATAAAC ATCGTGTTAC TTTCTAAAAA CCTTCCATTA AAAATATGTT TCTCGAGTTA
 Zk899    GTTTCAGGAC GAAGTTAGTA TGTTCAAACA AT..AAAAAA TTATATAAAC ATCGTGTTAC TTTCTAAAAA CCTTCCATTA AAAATATGTT TCTCGAGTTA
D1009rc   GTTTCAGGAC GAAGTTAGTA TGTTCAAACA AT..AAAAAA TTATATAAAC ATCGTGTTAC TTTCTAAAAA CCTTCCATTA AAAATATGTT TCTCGAGTTA
 M02d8    GTTTCAGGAC GAAGTTAGTA TGTTCAAACA AT..AAAAAA TTATATAAAC ATCGTGTTAC TTTCTAAAAA CCTTCCATTA AAAATATGTT TCTCGAGTTA
F02d10    GTTTCAGGAC GAAGTTAGTA TGTTCAAACA AT..AAAAAA TTATATAAAC ATCGTGTTAC TTTCTAAAAA CCTTCCATTA AAAATATGTT TCTCGAGTTA
 F19h6    GTTTCAGGAC GAAGTTAGTA TGTTCAAACA AT..AAAAAA TTATATAAAC ATCGTGTTAC TTTCTAAAAA CCTTCCATTA AAAATATGTT TCTCGAGTTA
 C30g4    GTTTCAGGAT AAAGTTAATA TGTGCAAACA GTCAAAAAAT ATATATAAAC ACCGTATTAC TTTC.ACAAA CCTTCCATTA AAAATATAGC TCTTGAGTTA

          701                                                                                                           800
C07d10    CAGTCGTTTT AGTAATCATC ATCAAAAGCG GACTCACGAG ATGTCGGCCG CTTTATCACT TGCCATGTTT TGGGTGAATA ATTTAATCAT GTAAAAATAC
 Zk899    CAGTCGTTTT AGTAATCATC ATCAAAAGCG GACTCACGAG ATGTCGGCCG CTTTATCACT TGCCATGTTT TGGGTGAATA ATTTAATCAT GTAAAAATAC
D1009rc   CAGTCGTTTT AGTAATCATC ATCAAAAGCG GACTCACGAG ATGTCGGCCG CTTTATCACT TGCCATGTTT TGGGTGAATA ATTTAATCAT GTAAAAATAC
 M02d8    CAGTCGTTTT AGTAATCATC ATCAAAAGCG GACTCACGAG ATGTCGGCCG CTTTATCACT TGCCATGTTT TGGGTGAATA ATTTAATCAT GTAAAAATAC
F02d10    CAGTCGTTTT AGTAATCATC ATCAAAAGCG GACTCACGAG ATGTCGGCCG CTTTATCACT TGCCATGTTT TGGGTGAATA ATTTAATCAT GTAAAAATAC
 F19h6    CAGTCGTTTT AGTAATCATC ATCAAAAGCG GACTCACGAG ATGTCGGCCG CTTTATCACT TGCCATGTTT TGGGTGAATA ATTTAATCAT ATAAAAATAC
 C30g4    CAGTCGTTTT AGTAACCATC ATCAAAAATG GACTCACGAG ATGTCGGCCG CTTTATTACT TGTCACGATT TTATTG.ATT TTTCAAACAT GTAAAAACAT
```

```
         801                                                                                                                  900
C07d10   ATAGGCCTGA TCACGAGCTA CATTTTACCA GTTGAAAC.T TTTTTGATAG CTTTCCTAGT TTTGGAGTTA TGCTCAATGG AAAAAATATC GAAATCATCA
Zk899    ATAGGCCTGA TCACGAGCTA CATTTTACCA GTTGAAAC.T TTTTTGATAG CTTTCCTAGT TTTGGAGTTA TGCTCAATGG AAAAAATATC GAAATCATCA
D1009rc  ATAGGCCTGA TCACGAGCTA CATTTTACCA GTTGAAACTT TTTTTGATAG CTTTCCTAGT TTTGGAGTTA TGCTCAATGG AAAAAATATC GAAATCATCA
M02d8    ATAGGCCTGA TCACGAGCTA CATTTTACCA GTTGAAACTT TTTTTGATAG CTTTCCTAGT TTTGGAGTTA TGCTCAATGG AAAAAATATC GAAATCATCA
F02d10   ATAGGCCTGA TCACGAGCTA CATTTTACCA GTTGAAAC.T TTTTTGATAG CTTTCCTAGT TTTGGAGTTA TGCTCAATGG AAAAAATATC GAAATCATCA
F19h6    ATAGGCCTGA TCACGAGCTA CATTTTACCA GTTGAAAC.T TTTTTGATAG CTTTCCTAGT TTTGGAGTTA TGCTCAATGG AAAAAATATC GAAATCATCA
C30g4    ATTTACCTGA TCACGGGCTA TACATTACCA GTTGAAA.AT TTTTTGATAG GATTTTTAGT TTTTGAGTTA TACTTATTGG AAAAAATAAC TAAATCAT.A

         901                                        946
C07d10   AAAAACAGAA AGTGGATATC TTTTTGGCCA GCACTGTATT TCGACC
Zk899    AAAAACAGAA AGTGGATATC TTTTTGGCCA GCACTGTAGT TTTTTC
D1009rc  AAAAACAGAA AGTGGATATC TTTTTGGCCA GCACTGTACG TGACTA
M02d8    AAAAACAGAA AGTGGATATC TTTTTGGCCA GCACTGTATA TACGTG
F02d10   AAAAACAGAA AGTGGATATC TTTTTGGCCA GCACTGTATA TACACA
F19h6    AAAAACAGAA AGTGGATATC TTTTTGGCCA GCACTGTACA TTACAT
C30g4    AAAAACTAAA AGTGGATATC TTTTTGGCCA GCACTGTATA TATAGT
```

APPENDIX C: Alignment of a modified Tc2 sequence and ten cosmid sequences identified as high scoring blast hits to Tc2.

```
         1                                                                                                  100
F56d5rc  ACAAAACATA CCGTATATTC TCTATTAGTG CTGCATATCC AGTAGTACTG CAGTCTCTAA TAGTGCTGCA TTCAAAAAAT ..GTCCAAAA AATAGTGCTG
T26a8rc  AATATTATTA CCGTATATTC TCTATTAGTG CTGCATATCC AGTAGTACTG CAGTCTCTAA TAGTGCTGCA TTCAAAAAGT ..GTCCAAAA AATAGTGCTG
  T10f2  GGTTTATATA CCGTATATTC TCTATTAGTG CTGCATCTCC AGTAGTACTG CAGTCTCTAA TAGTGCTGCA TTCAAAAAGG ..GTCCGAAA AATAGTGCTG
  Zk792  TTCAAACATA CCGTATATTC TCTATTAGTG CTGTATCTCC AGTAGTACTG CAGTCTCTAA TAGTGCTGCA TTCAAAAAGG ..GTCCGAAA AATAGTGCTG
F53h8rc  CAATCATTTA CCGTATATTC TCTATTAGTG CGGCATCTCT A......... .......... .......... ......CTGC CCTTTCGAAA ATTAGTGCGG
Zk455rc  TTTTAAAGTA CCGTATATTC TCTATTAGTG CGGCATCTCT AATAGTGCGG C.ATCTCTAT TAGTGCGGCG CCCCTACTGC CCTTTCGAAA ATTAGGGCGG
F01f1rc  TATATATATA CCGTATATTC TCTATTAGAG CGGCGTATCT AATAGTACGG C.GTATCTAT TAGAGCGGCA CCCTAACTGG CTCTTCGAAA ATTAGAGCGG
 F52b10  ATTTAATATA CCGTATATTC TCTATTAGAG CGGCGTATCT AATAGTACGG C.GTATCTAT TAGAGCGGCA CTCTAACTGG CTCTTCGAAA ATTAGAGCGG
K03h9rc  TTAAGAGATA CCGTATATTC TCTATTATAA CGGCGTATCT AATAGAACGG G.GTATCTAT TAGAGCGGCA CCCTAACTGG CGCTTCGAAA ATTAGAGCGG
 C15b12  AAAAATATTA CTGTATACTC TCTATTAGTG CGGCGTAAGT AATACTACAG C.GTATTTAT TAGTGAGGCA CTTTAACTGG CCTTCTAAAA AATAATACGG
 Tc2del  .......... CCGTATATTC TCTATTAGTG CTGCGTATCT AATAGAACGG C.GTATCTAA TAGAGCTGCA CCCAACAGGC ATTTCCGAAA ATTACAGCGG

         101                                                                                                 200
F56d5rc  CAGTCTCTAT TAGTGCTGCA .......... .......... .......... ......CCCC ATTTTCTATT GGAAGAGTCA CTTGACAGTT TCGATATTTT
T26a8rc  CAGTCTCTAT TAGTGCTGCA .......... .......... .......... ......CCCC ATTTTCTATT GAAAGAGTCA CTTGATATTT TCGATATTTT
  T10f2  CAGTCTCTAT TAGTGCTGCA .......... .......... .......... ......CCCC ATTTTCTATT GAAAGAGTCA CTTGACAGTT TCGATATTTT
  Zk792  CAGTCTCTAT TAGTGCTGCA .......... .......... .......... ......CCCC ATTTTCTATT GAAAGAGTTA CTTGACAGTT TCGATATTTT
F53h8rc  CAGTCTCCAG TAGTCCGGCA GTCTCTAATA GTGCGGCAC. .......... .....CCCTT CTTTTTGGTT GCGAAAGCCA CCTCAAATAT TCGGATTCTA
Zk455rc  CAGTCTCTAA TAGTCCGGCT GTCTCTAATG GTGCGGCAC. .......... .....CCCTT CTTTTTGGTT GCGAAAGCCA CATCAAATAT TCGATTACTC
F01f1rc  CAGTCTCTAT TAGTGTGGCA .......... .......... .......... .....CCCCT TTTTTTGGGT GCGAGAGCCA CCTCAATTAT TCGATTTCTT
 F52b10  CAGTCTCTAT TAGTGCGGCA .......... .......... .......... .....CCCCT TTTTTTGGGT GCAAGAGCCA CATCAATTAT TCGATTTCTT
K03h9rc  CAGTCTTTAA TAGTGCGGCA .......... .......... .......... .....CCCTT TTTTTTAGAT GCGAAAGCCA CCTCAATTAT TCGATTACTT
 C15b12  CAGTCCCTAT TAGTGCGGCA .......... .......... .......... .....CACCA TTTGTGGGTT GCGAAAGCCA TTACAAATAT TCGATTACTT
 Tc2del  CAGTCTCCAA TAGAGCGGCA GTCTCTAATA GAGCGGCACC CTTTTCGCTG CGAGACCCGC CGCTGCTTTT CAGCTTATTT TTTGAATTTT TTTCAGTTTA
```

```
         201                                                                                                      300
F56d5rc  TGGG.TATTT TTTTTGTATA TTTTTGATGA TTTTTTGATA ATTTACAAAA GTAATTGAAT TATTGGGTCT TTTTGTTACA AAAGTTCGCA ATTTTTATGA
T26a8rc  TGGGTTGTTT TTTTTGTATA TTTTTGATGA TTTTTTGATA ATTTACGAAA GTAATTGAAT TATTGGGTCT TTTTGTTACA AAAGTTCGCA ATTTTTATGA
   T10f2 TGGG....TA TTTTTTTGTA TATTTTATGA TTTTTTGATA ATTTAGGAAA GTAATTGAAT TATTGGGTCT TTTTTTTACA AAAGTTCGCA ATTTTTATGA
   Zk792 TGG....TTT TTTTTGTATA TTTTTGATGA TTTTTTGATA AT......AA GTAATTGAAT TATTGGGTTT TTTTGTTACA AAAGTTCGCA ATTTTTATGA
F53h8rc  CAATGGTATC CTTGTTCATT CTTCAATAAT TTCTCATAAA TTTTCACGTA AACATTGCAT TTCAGAGTCT TT.GGTGGCA AAA.TACACA ATTTTCATTA
Zk455rc  CAATGGTATT CTTGTTCATT CTTCAATAAT TTCTCATAAA TTTTCACTTT ATCATTGCAT TTCAGTTTCT TTGGGGGGCA AAA.TACACA ATTTTCATTA
F01f1rc  CGAAGATATT TTTGTTCATT CTTGAATGAA TTC..ATAAA TTTTCATTTA AATATTGAAT TTTAGTGTCT TTGTGAGGCA AAATTACACA ATTTTCACGA
  F52b10 CGAAGATATT TTTGTTCATT CTTGAATGAA TTT..ATAAA TTTTCATTTA AATATTGAAT TTTAGTGTCT TTGTGAGGCA AAATTACCCA ATTTTCATGA
K03h9rc  CGATGGTATT TTTGTTTATT CTTCAATGAA TTC..ATAAA TTTTCATTTA AACATTGAAT TTTAGTGTCT TTGCGAGGCA AAAATACACA ATTTTAATGA
  C15b12 CGATGGTTTT CTTGTTTATT CTTCAATGGT TTCCTATAAA TTTTCTCTTT AACGTTGACT TTTAGTGTCT TCGTGAGGCA GAATTACACA ATTTCAATGA
  Tc2del TTTTCTTTTT TTTCAACATT TTTCAATTGT TTTACGTGGT TTTTTTGTTT GAAATCAATT CTATCATTTC CGTAGATAAT TTTTGTTTTT TTTCTGTTGT

         301                                                                                                      400
F56d5rc  ATTTCTGCTT AAAAAATGAG CTTTTGTGA. .......... .......... CTAAAAACTA GTGA.....T CTATAATTGA AAATAGCCGT TCGAAAATTA
T26a8rc  ATTTCTGCTT AAAAAATGAG CTTTTGTGA. .......... .......... CTGAAAACTA GTGA.....T CTAAAATTGA AAATAGCCGT TCGAAAATTA
   T10f2 ATTTCTGCTT AAAATATGAG GTTTTGTGA. .......... .......... TTAAAAACTA GTGA.....T CTATAATTGA AAATAGCCGT TCGAAAATTA
   Zk792 ATTTCTGCTT AAAA.ATGAG TTTTTGGGA. .......... .......... CTAAAAACTA GTGA.....T CTATAATTGA AAATAGCCGT TCGAAAATTA
F53h8rc  CTTTTCACAT AAAATATGAG CTTTTATGAG GGAATTTTGA GGTTACTACT CTCGAAAGTA GTAA.....C CTCAAAGTGA AAATTAGTTA TCGAAAATTA
Zk455rc  ATTTTCACAT AACATATGAG GTTTTATGAG GGAATTTTGA GTTTACTACT CTCGAAAGTA GTTA.....C CTCAAAATGA AAATTAGTTA TCGAAAATTA
F01f1rc  ATTTTTCCAT AAAATATGAG GTTTTATAAG GAAATTTTGA GGTTACTACT CCCAACAGTA GTAA.....C TTCAATATGA AAATTAGTTA TCGAAAATTA
  F52b10 ATTTTTCCAT AAAATATGAG GTTTTATAAA GAAATTTTGA GGTTACTACT CCCAAAAGTA GTAA.....C TTCAATATGA AAATTAGTTA TCGAAAATTA
K03h9rc  A.TTTCACAT AAAATATGAG CTTTTATAAG GAAATTTTGA GGTTTCTACT CTCAAAAGTA TTAA.....T TTAAAAATAA AAATTAGTTA TCGAAAATTA
  C15b12 ATTTTCAAAT AAAGTATGAG GTTTTATAAG AGAATTTTGA ..TTACTAAT TTTGAAGGTA GTAA.....C CTCAAAATAA AAATTA.TTA TCGAAAATTA
  Tc2del TTGTTTTTCT TTGATTTTTT TTTCATTTAT GAAAAATTGT TATTTCTACT CCACCATTAA TTGAAACTCC TGAAAACAAA AAAAACTCGT TCGAAAATTA
```

```
           401                                                                                                       500
F56d5rc    ATGCTGCAGT CTCCAGTAGT ACTGCAGTCT CTAATAGTGC TGCATTAAAA GAGTGCCTGG AAATTAGTGC TGCATGCAGC ACTAATAGAG AATATACGGT
T26a8rc    ATGCTGCAGT CTCCAGTAGT ACTGCAGTCT CTAATAGTGC TGCATTAAAA AAGGGCCTGG AAATTAGTGC TGCATGCAGC ACTAATAGAG AATATACGGT
  T10f2    ATGCTGCAGT CTCCAGTAGT ACTGCAGTCT CTAATAGTGC TGCATTAAAA AAGGGCCTGG AAATTAGTGC TGCATGCAGC ACTAATAGAG AATATACGGT
  Zk792    ATGCTGCAGT CTCCAGTAGT ACTGCAGTCT CTAATAGTGC TGCATTAAAA AAGGGCCT.G AAATTAGTGA TGCATGCAGC ACTAATAGAG AATATACGGT
F53h8rc    GTACGGCAGT CTCCAGTAGT CCGGCAGTCT CTAATAGTGC GCAAGTCGTG GAAGGCTCGG AAATTAGTGC GGCATGCCGC ACTAATAGAG AATATACGGT
Zk455rc    GTACGGCAGT CTCCAGTAGC CCGGCAGTCT CTAATAGTGC GCAAGTCACG AAAGGCCCGG AAATTAGTGC GGCATGCCGC ACTAATAGAG AATATACGGT
F01f1rc    GTGCGGCAGT ATCCAGTAGA ACGGCAGTCT CTAATAGAGC GCACGTCGCG AAGGGCCCGG AAATTAGTGC GGCATGCCGC TCTAATAGAG AATATACGGT
 F52b10    GTGCGGCAGT ATCCAGTAGA ACGGCAGTCT CTAATAGAGC GCACGTCACG AAGGGCCGGG AAATTAGTGC GGCATGCCGC TCTAATAGAG AATAGAGAAT
K03h9rc    GTACGGCAGT CTCCAGTAGA AGGGCAGTCT CTAATAGTGG TATGTATTTT CCCAATGAGG GAGGCAATGT TAATGGCAGA TAATAGCAAA G.........
 C15b12    GTGCGGCAGT GTCCAGTAGA ACGGCAGTCT GTAATAGTGC ACAAGTCACG GAAGGGCCGG AAATTAGTGA GGCATGCCGC ACTATTTGAA AAAAAAAGTC
 Tc2del    GAGCGGCAGT CTCCAGTAGA ACGGCAGTCT CTAATAGAGC TGCAATCGAG AAGGCTCTGA AAATTAGAGC GGCATGCAGC ACTAATAGAG AATATACGG.

           501
F56d5rc    ACCTTTTAA
T26a8rc    ATTATGAAA
  T10f2    ATTTGCTGA
  Zk792    AACTTGGGT
F53h8rc    AATTCTTTA
Zk455rc    AAATACGTT
F01f1rc    ATATAACGT
 F52b10    ACACGGTAC
K03h9rc    .........
 C15b12    AATTCGGAG
 Tc2del    .........
```

APPENDIX D: Alignment of Tc3 and ten cosmid sequences identified as high scoring blast hits to Tc3.

```
          1                                                                                                         100
   F27E5  CAGTGTGAGA AAGTTCTATA GGACCCCCCT CATTTTTGAC GTTTCACCTA TATTGAAAAT TTTTTTTGAA AAATTTTTTC CGTAGATTGA TTCAAAATGT
  T13A10  CAGTGTGAGA AAGTTCTATA GGACCCCCCT CATTTTTGAC GCTTCCCCTA TTTTGAAAAT TTTTTTTTGA AAAATTTTAC CGTGGATTGA TTTAAAATGT
   K10B3  CAGTGTGGGA AAGTTCTATA GGACCCCCCC TAATTTGAAG GTTTGAGGAA CTTCCGAAAA TTTTTTTGAA AAACTGCTAA TGCCATTCGT TTTTAAATTG
ZK1086RC  CAGTGTGGGA AAGTTCTATA GGACCCCCCC TAATTTGAAG GTTTGAGGAA CTTCCGAAAA TTTTTTTGAA AAACTGCTAA TGCCATTCGT TTTTAAATTG
 R10H1RC  CAGTGTGGGA AAGTTCTATA GGACCCCCCC TAATTTGAAG GTTTGAGGAA CTTCCGAAAA TTTTTTCGAA AAACTGCTAA TGCCGTTCGT TTTTAAATTG
   T02G5  CAGTGTGGGA AAGTTCTATA GGACCCCCCC TAATTTGAAG GTTTGAGGAA CTTCCGAAAA TTTTTTCGAA AAACTGCTAA TGCCGTTCGT TTTTAAATTG
   B0303  CAGTGTGGGA AAGTTCTATA GGACCCCCCC TAATTTGAAG GTTTGAGGAA CTTCCGAAAA TTTTTTCGAA AAACTGCTAA TGCCGTTCGT TTTTAAATTG
     Tc3  CAGTGTGGGA AAGTTCTATA GGACCCCCCC TAATTTGAAG GTTTGAGGAA CTTCCGAAAA TTTTTTCGAA AAACTGCTAA TGCCGTTCGT TTTTAAATTG
T25G12RC  CAGTGCGCCC AACTTCTATA GCGCCCCCCT ATTTTTTCTC ATTTCACCAC CTTCTGACAT TTTTTACAAA AAACTGCTAA TGCCATTCGA TTCCAAATGT
    ZC64  CAGTGCGCCC AACTTCTATA GCGCCCCCCT ATTTTTTCTC ATTTCACCAC CTTCTGACAT TTTTTACAAA AAACTGCTAA TGCCATTTGA TTCCAAATGT
   C25G4  CAGTGCGCCC AACTTCTATA ACGCCCCCCT ATTTTTTCTC ATTTCACCAC CTGCTGACAT TTTTTACAAA AAACTGCTAA TGCCATTCGA TTCCAAATGT

          101                                                                                                       200
   F27E5  CAAAATA... TGTGGACACA AAAATTTTCA GACTTGTACC T...CTAAAA CTG.GCTGCG TACGGAAAAA AGTTGGAAAA TCAGGTGTGC AACTTCTATA
  T13A10  CAAAATA... TGTGGGCACA AAAATTTTCA GAGGTGTACC T...CTAAAA CTG.GCTGCG TACCAAAAAA TGTTGGAAAA TAAGGTGTGC AACTTCTATA
   K10B3  AAAAAAA... CCTATATACA TTTTTTTCCA GAAGTTTATC T...CAAAAA CTGAGGTCGC GCTG.GAAAA AACGTCAAAA TCCAGTGTGA AACTTCTATA
ZK1086RC  AAAAAAA... CCTATATACA TTTTTTTCCA GAAGTTTATC T...CAAAAA CTGAGGTCGC GCTG.GAAAA AACGTCAAAA TCCAGTGTGA AACTTCTATA
 R10H1RC  AAAAAAA... CCTATATACA TTTTTTTCCA GAAGTTTATC T...CAAAAA CTGAGGTCGC GCTG.GAAAA AACGTCAAAA TCCAGTGTGA AACTTCTATA
   T02G5  AAAAAAA... CCTATATAAA TTTTTTTCCA GAAGTTTATC T...CAAAAA CTGAGGTCGC GCTG.GAAAA AACGTCAAAA TCCAGTGTGA AACTTCTATA
   B0303  AAAAAAA... CCTATATACA TTTTTTTCCA GAAGTTTATC T...CAAAAA CTGAGGTCGC GCTG.GAAAA AACGTCAAAA TCCAGTGTGA AACTTCTATA
     Tc3  AAAAAAA... CCTATATACA TTTTTTTCCA GAAGTTTATC T...CAAAAA CTGAGGTCGC GCTG.GAAAA AACGTCAAAA TCCAGTGTGA AACTTCTATA
T25G12RC  CGAAAAACCC CCTGTCCATA ATTTTCATCA AAAAATTCAA TAGAC..... .TGAGGAGAC TAGGTCAAAA AGTCTCTAAA ACGACCGCCC AACTTCTATA
    ZC64  CGAAAAACCC CCTGTCCAAA ATTTTCATCA AAAAATTCAA TAGAC..... .TGAGGAGAC TAGGTCAAAA AGTCTCTAAA ACGACCGCCC AACTTCTATA
   C25G4  CGAAAAACCC CCTGTCCAAA ATTTTCATCA AAAAATTCAA TAGACTGATA GTGAGGAGAC TAGGTCAAAA AGTCTCTAAA ACGACCGCCC AACTTCTATA
```

```
          201                                                                                                          300
   F27E5  GGACCCCCTC ATATTTTTGA CGATTTCACC GGAAAAAAAT AGTTTTTGAA GTT....... AATATCACGT AAAT....AG TTATGAAAAT GAGAATAAAC
  T13A10  GGACCCCCTT ATATTTTTGA CGATTTCATT GAGAAAAAAG CGTTTTTGGA ATG....... AATATCCCGT AAAT....TA TTATGAAAAT GAGAATAAAC
   K10B3  GGACCCCCCG TTTTTTTTCA CGATTTTTAC TAAAATCAAC AGATTTTGGA ATTTTTGACA AAGCTCAAAT CAAGTTTGAG TTAGAAATGA GTTCATATAA
ZK1086RC  GGACCCCCCG TTTTTTTTCA CGATTTTTAC TAAAATCAAC AGATTTTGGA ATTTTTGACA AAGCTCAAAT CAAGTTTGAG TTAGAAATGA GTTCATATAA
 R10H1RC  GGACCCCCCG TTTTTTTTCA CGATTTTTAC TAAAATCAAC AGATTTTGGA ATTTTTGACA AAGCTCAAAT CAAGTTTGAG TTAGAAATGA GTTCAGATAA
   T02G5  GGACCCCCCG TTTTTTTTCA CGATTTTTAC TAAAATCAAC AGATTTTGGA ATTTTTGACA AAGCTCAAAT CAAGTTTGAG TTAGAAATGA GTTCATATAA
   B0303  GGACCCCCCG TTTTTTTTCA CGATTTTTAC TAAAATCAAC AGATTTTGGA ATTTTTGACA AAGCTCAAAT CAAGTTTGAG TTAGAAATGA GTTCAGATAA
     Tc3  GGACCCCCCG TTTTTTTTCA CGATTTTTAC TAAAATCAAC AGATTTTGGA ATTTTTGACA AAGCTCAAAT CAAGTTTGAG TTAGAAATGA GTTCATATAA
T25G12RC  GCGCCCCCTC GAATTTTTGA TTTTTTCTAA TAAAATCCCC TTTTTCTGGC TTTTT...CT CAGAACAACT TATTCATGAC TTAAGTTCAA ACTGATCCAG
    ZC64  GCGCCCCCTC GAATTTTTGA TTTTTTCTAA TAAAATCCCC TTTTTTTGGC TTTTT...CT CAGAACAACT TATTCATGAC TTAAGTTCAA ACTGATCCAG
   C25G4  GCGCCCCCTC GAATTTTTTA TTTTTTCTAA TAAAATCCCC TTTTTTTGGC TTTTT...CT AAGAACAACT TATTCATGAC TTAAGTTCAA ACTGATCCAG

          301                                                                                                          400
   F27E5  ACAGTTT... .......... ....CAAGCA ATTTCATATT GGCACTTGGT GGTTGACCCA CAACTGTCCT CTAACAGTCC TGTAACAGAA GTTTCCACGT
  T13A10  ACAGTTT... .......... ....CAAGCA ATTTCCTATT GGCATTTGAG GGTTGACCCG CAACTGTCCT GTAACAGTCA TGTAACAGAA GTTTCCACGT
   K10B3  GCAGTTTTGA CTTT.AAAAA TTAATACGAA ATGTTCTCGT GGGATCTCCA GACTGGTTCT GATTCTTCCG ATCTTTGATG TTCAAGTCTG TTTCAAGCTT
ZK1086RC  GCAGTTTTGA CTTT.AAAAA TTAATACGAA ATGTTCTCGT GGGATCTCCA GACTGGTTCT GATTCTTCCG ATCTTTGATG TTCAAGTCTG TTTCAAGCTT
 R10H1RC  GCAGTTTTGA CTTTAAAAAA TTAATACGAA ATGTTCTCGT GGGATCTCCA GACTGGTTCT GATTCTTCCG ATCTTTGATG TTCAAGTCTG TTTCAAGCTT
   T02G5  GCAGTTTTGA CTTTAAAAAA TTAATACGAA ATGTTCTCGT GGGATCTCCA GACTGGTTCT GATTCTTCCG ATCTTTGATG TTCAAGTCTG TTTCAAGCTT
   B0303  GCAGTTTTGA CTTT.AAAAA TTAATACGAA ATGTTCTCGT GGGATCTCCA GACTGGTTCT GATTCTTCCG ATCTTTGATG TTCAAGTCTG TTTCAAGCTT
     Tc3  GCAGTTTTGA CTTT.AAAAA TTAATACGAA ATGTTCTCGT GGGATCTCCA GACTGGTTCT GATTCTTCCG ATCTTTGATG TTCAAGTCTG TTTCAAGCTT
T25G12RC  ACATCTTGCA AAT..CAACT TTCAAAAGAA ATTTATCCGT GGAA...... .......... .ATTCGTAGG ATGTCT.... .......... ..........
    ZC64  ACATCTTGCA AAT..CAACT TTCAAAAGAA ATTTATCCGT GGAA...... .......... .ATTCGTAGG ATCTCT.... .......... ..........
   C25G4  ACATCTTGCA AAT..CAATT TTCAAAAGAA ATTTATCCGT GGAA...... .......... .ATTCATAGG ATCTCT.... .......... ..........
```

```
          401                                                                                                       500
   F27E5  GTCATCTGCC CTATCACTAC CAAACTTTAC AAAATGATAC ATGACACTGT TCTAACAACA TCCCTA.... .......... .......... ..........
  T13A10  GTCACTTACC CTATCACAAC CACACTTTAT AAAATGAGAC ATGACACTGT TCTAACAACA TTCGTA.... .......... .......... ..........
   K10B3  CCTGGTGCTC TCGGTAATGC CAAAACTTGA TAAACTCTCT TTAACAAGTT CCTACTAAAA TTCCTAGCAC GCACTTTAGA TGTTTCGACT GTGTAGTCAA
ZK1086RC  CCTGGTGCTC TCGGTAATGC CAAAACTTGA TAAACTCTCT TTAACAAGTT CCTACTAAAA TTCCTAGCAC GCACTTTAGA TGTTTCGACT GTGTAGTCAA
 R10H1RC  CCTGGTGCTC TCGGTAATGC CAAAACTTGA TAAACTCTCT TTAACAAGTT CCTACTAAAA TTCCTAGCAC ACACTTTAGA TGTTTCGACT GTGTAGTCAA
   T02G5  CCTGGTGCTC TCGGTAATGC CAAAACTTGA TAAACTCTCT TTAACAAGTT CCTACTAAAA TTCCTAGCAC GCACTTTAGA TGTTTCGACT GTGTAGTCAA
   B0303  CCTGGTGCTC TCGGTAATGC AAAAACTTGA TAAACTCTCT TTAACAAGTT CCTACTAAAA TTCCTAGCAC GCACTTTAGA TGTTTCGACT GTGTAGTCAA
     Tc3  CCTGGTGCTC TCGGTAATGC CAAAACTTGA TAAACTCTCT TTAACAAGTT CCTACTAAAA TTCCTAGCAC GCACTTTAGA TGTTTCGACT GTGTAGTCAA
T25G12RC  .......... .......... .......... .......... .....AAAAT TGTCCTGAAA CGCCTA.... ..ATTTTA.. .......... ..........
    ZC64  .......... .......... .......... .......... .....AAAAT TGTCCTGAAA CGCCTA.... ..ATTTTA.. .......... ..........
   C25G4  .......... .......... .......... .......... .....AAAAT TGTCCTGAAA CGCCTA.... ..ATTTTA.. .......... ..........

          501                                                                                                       600
   F27E5  ..AGTTTTTG GAAAA..... ATCCGAAAAG TTTGAGCCCC GCTGCACCGC TGAAACCG.. .......... ........TA CATGGGCCTC ATTTTTCA.G
  T13A10  ..CGTTTTTG GAAAA..... ATCCGAAAAG TTTGAGCTCC GCTGCACCGC TGAAACCG.. .......... ........TA CATGGGCCTA TTATTTCA.G
   K10B3  GCTGATTTGG CAAAATATGC AGCAGGAAAC AATGGAAGGC TTATCAGGAA TCAAATCGTT TTTCTTTGAT TACAAGGTTC CATGGGACCA ATATTTCAAG
ZK1086RC  GCTGATTTGG CAAAATATGC AGCAGGAAAC AATGGAAGGC TTATCAGGAA TCAAATCGTT TTTCTTTGAT TACAAGGTTC CATGGGACCA ATATTTCAAG
 R10H1RC  GCTGATTTGG CAAAATATGC AGCAGGAAAC AATGGAAGGC TTATCAGGAA TCAAATCGTT TTTCTTTGAT TACAAGGTTC CATGGGACCA ATATTTCAAG
   T02G5  GCTGATTTGG CAAAATATGC AGCAGGAAAC AATGGAAGGC TTATCAGGAA TCAAATCGTT TTTCTTTGAT TACAAGGTTC CATGGGACCA ATATTTCAAG
   B0303  GCTGATTTGG CAAAATATGC AGCAGGAAAC AATGGAAGGC TTATCAGGAA TCAAATCGTT TTTCTTTGAT TACAAGGTTC CATGGGACCA ATATTTCAAG
     Tc3  GCTGATTTGG CAAAATATGC AGCAGGAAAC AATGGAAGGC TTATCAGGAA TCAAATCGTT TTTCTTTGAT TACAAGGTTC CATGGGACCA ATATTTCAAG
T25G12RC  ..TGTTAGAG GGCGCTGCGC GGCGCCCAAT ACTTGAATCA GCACCA.GAA TCAACT.... .......... ...AAGATGA TCTACAACAT CTATTTCA..
    ZC64  ..TGTTAGAG GGCGCTGCGC GGCGCCCAAT ACTTGAATCA GCACCA.GAA TCAACT.... .......... ...AAGATGA TCTACAACAT CTATTTCA..
   C25G4  ..TGTTAGAG GGCGTTGCGC GGCGCCCAAT ACTTGAATCA GCACCA.GAA TCAACT.... .......... ...AAGATGA TCTACAACAT CTATTTCA..
```

```
          601                                                                                                                  700
   F27E5  GCCGTTTCTC CGTTTCAGTG TGTATAACTC GGGT...... .......... .......... .......... .......... .......CAC GGTGAACCCC
  T13A10  GCCATTTCTC CGT....... TGTACTAATC CGGT...... .......... .......... .......... .......... .......CCC GGTGAGCCCC
   K10B3  TTAAATTGTC CCTCACAGAT GTTATTTACT ATTTTTTGCG TGAATTATTA AATGTGGAAT TGTGGCATGT GTTGTGGCAC ACATATAGAG GCTGGAAAGC
ZK1086RC  TTAAATTGTC CCTCACAGAT GTTATTTACT ATTTTTTGCG TGAATTATTA AATGTGGAAT TGTGGCATGT GTTGTGGCAC ACATATAGAG GCTGGAAAGC
 R10H1RC  TTAAATTGTC CCTCACAGAT GTTATTTACT ATTTTTTGCG TGAATTATTA AATGTGGAAT TGTGGCATGT GTTGTGGCAC ACATATAGAG GCTGGAAAGC
   T02G5  TTAAATTGTC CCTCACAGAT GTTATTTACT ATTTTTTGCG TGAATTATTA AATGTGGAAT TGTGGCATGT GTTGTGGCAC ACATATAGAG GCTGGAAAGC
   B0303  TTAAATTGTC CCTCACAGAT GTTATTTACT ATTTTTTGCG TGAATTATTA AATGTGGAAT TGTGGCATGT GTTGTGGCAC ACATATAGAG GCTGGAAAGC
     Tc3  TTAAATTGTC CCTCACAGAT GTTATTTACT ATTTTTTGCG TGAATTATTA AATGTGGAAT TGTGGCATGT GTTGTGGCAC ACATATAGAG GCTGGAAAGC
T25G12RC  ........TC GGTCTGAAA. TTTTTTTCAC GTC.GACGTG TCA....... .......... .........T TTTGTGGCAT ACGAATGCTC AAGAAAACGT
    ZC64  ........TC GGTCTGAAA. TTTTTTTCAC GTCTTACGTG TCA....... .......... .........T TTTGTGGCAT ACGAATGCTC AAGAAAACGT
   C25G4  ........TC GGTCTGAAAT TTTTTTTCAC GTCTTACGTG TCA....... .......... .........T TTTGTGGCAT ACGAATGCTC AAGAAAACGT

          701                                                                                                                  800
   F27E5  CGACTTTGAA AACTGTTCAA .......... .......... TAAACTCGTA TGAACATTAC GCTAGCATCC ACA....... .......... ..........
  T13A10  TAACTTTGAA AATTGTTGAA .......... .......... TAAACTTGCA TGAACATCAA GCTAGCATCC ACA....... .......... ..........
   K10B3  TTACTTCGAA AGCAGTCTAA CTTGCAATGC CTCGAGGATC TGCCCTTTCG GACACTGAAC GCGCTCAGCT GGATGTTATG AAATTGCTCA ATGTGTCCCT
ZK1086RC  TTACTTCGAA AGCAGTCTAA CTTGCAATGC CTCGAGGATC .......... .......... .......... .......... .......... ..........
 R10H1RC  TTACTTCGAA AGCAGTCTAA CTTGCAATGC CTCGAGGATC TGCCCTTTCG GACACTGAAC GCGCTCAGCT GGATGTTATG AAATTGCTCA ATGTGTCCCT
   T02G5  TTACTTCGAA AGCAGTCTAA CTTGCAATGC CTCGAGGATC TGCCCTTTCG GACACTGAAC GCGCTCAGCT GGATGTTATG AAATTGCTCA ATGTGTCCCT
   B0303  TTACTTCGAA AGCAGTCTAA CTTGCAATGC CTCGAGGATC TGCCCTTTCG GACACTGAAC GCGCTCAGCT GGATGTTATG AAATTGCTCA ATGTGTCCCT
     Tc3  TTACTTCGAA AGCAGTCTAA CTTGCAATGC CTCGAGGATC TGCCCTTTCG GACACTGAAC GCGCTCAGCT GGATGTTATG AAATTGCTCA ATGTGTCCCT
T25G12RC  AGTGTTTATT ATCAATTTTG TTCGAAATGC CACGAGGACC CCAGCTAACA AGAGATGAAC GATCCAAGCT AGATGTGATG GCGAATCTTA ATATTTCTAC
    ZC64  AGTGTTTATT ATCAATTTTG TTCGAAATGC CACGAGGACC CCAGCTAACA AGAGATGAAC GATCCAAGCT AGATGTGATG GCGAATCTTA ATATTTCTAC
   C25G4  AGTGTTTATT ATCAATTTTG TTCGAAATGC CACGAGGACC TCAGCTAACA AGAGATGAAC GATCCAAGCT AGATGTGATG GCGAATCTTA ATATTTCTAC
```

```
          801                                                                                                       900
    F27E5 .......... ......AAAA AA...TGGTC ACGATTTCGT GGTTGCAATG TTCT...... .......... .......... .......... ..........
   T13A10 .......... ......AAAA AAAATTGGTC ACGATTTTGT TATTGTAATG TTCT...... .......... .......... .......... ..........
    K10B3 GCATGAAATG AGTAGGAAAA TTTCCCGTTC TCGACACTGT A......... .......... .......... .......... .......... ..........
 ZK1086RC .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  R10H1RC GCATGAAATG AGTAGGAAAA TTTCCCGTTC TCGACACTGT ATTCGCGTGT ATCTGAAGGA TCCGGTGAGC TACGGTACAT CTAAAAGAGC TCCTCGTCGC
    T02G5 GCATGAAATG AGTAGGAAAA TTTCCCGTTC TCGACACTGT ATTCGCGAGT ATCTGAAGGA TCCGGTGAGC TACGGTACAT CTAAAAGAGC TCCTCGTCGC
    B0303 GCATGAAATG AGTAGGAAAA TTTCCCGTTC TCGACACTGT ATTCGCGAGT ATCTGAAGGA TCCGGTGAGC TACGGTACAT CTAAAAGAGC TCCTCGTCGC
      Tc3 GCATGAAATG AGTAGGAAAA TTTCCCGTTC TCGACACTGT ATTCGCGTGT ATCTGAAGGA TCCGGTGAGC TACGGTACAT CTAAAAGAGC TCCTCGTCGC
 T25G12RC AAATGAAATG GCTCGCCAAA TCAATCGCTC TCGTAAATGT GTCTACAACT ACCTCAATAG TCCACTTTCT TATGGTCAAA CCAAAAGAGC TCCCAGATGC
     ZC64 AAATGAAATG GCTCGCCAAA TCAATCGCTC TCGTAAATGT GTCTACAACT ACCTCAATAG TCCACTTTCT TATGGTCAAA CCAAAAGAGC TCCCAGATGC
    C25G4 AAATGAAATG GCTCGCCAAA TCAATCGCTC TCGTAAATGT GTCTACAACT ACCTCAATAA TCCACTTTCT TATGGTCAAA CCAAAAGAGC TCCCAGATGC

          901                                                                                                      1000
    F27E5 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
   T13A10 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
    K10B3 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
 ZK1086RC .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  R10H1RC AAAGCTCTCT CCGTGCGTGA CGAACGAAAT GTGATTCGTG CTGCCTCCAA CTCCTGTAAG ACGGCAAGAG ATATTCGCAA TGAGCTTCAA TTGTCTGCTT
    T02G5 AAAGCTCTCT CCGTGCGTGA CGAACGAAAT GTGATTCGTG CTGCCTCCAA CTCCTGTAAG ACGGCAAGAG ATATTCGCAA TGAGCTTCAA TTGTCTGCTT
    B0303 AAAGCTCTCT CCGTGCGTGA CGAACGAAAT GTGATTCGTG CTGCCTCCAA CTCCTGTAAG ACGGCAAGAG ATATTCGCAA TGAGCTTCAA TTGTCTGCTT
      Tc3 AAAGCTCTCT CCGTGCGTGA CGAACGAAAT GTGATTCGTG CTGCCTCCAA CTCCTGTAAG ACGGCAAGAG ATATTCGCAA TGAGCTTCAA TTGTCTGCTT
 T25G12RC AAAGTTTTAT CGAGTCGTGA GGAACGCAAC ATTGTGAAGG CTGCATCGAA CTCTTTCAAA TCTGCCAATG ATATTCGCAA GGAATTGAAT CTTAATGTTT
     ZC64 AAAGTTTTAT CGAGTCGTGA GGAACGCAAC ATTGTGAAGG CTGCATCGAA CTCTTTCAAA TCTGCCAATG ATATTCGCAA GGAATTGAAT CTTAATGTTT
    C25G4 AAAGTTTTAT CGAGTCGTGA GGAACGCAAC ATTGTGAAGG CTGCATCG.. .......... .......... .......... .......... ..........
```

```
          1001                                                                                                                  1100
   F27E5  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  T13A10  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
   K10B3  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
ZK1086RC  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
 R10H1RC  CAAAAAGGAC CATCCTCAAT GTCATCAAAC GATCTGGTGT AATCGTTCGT CAGAAACTTC GCCCTGCTCC GTTACTCTCT GCAGACCATA AACTCAAGCG
   T02G5  CAAAAAGGAC CATCCTCAAT GTCATCAAAC GATCTGGTGT AATCGTTCGT CAGAAACTTC GCCCTGCTCC GTTACTCTCT GCAGACCATA AACTCAAGCG
   B0303  CAAAAAGGAC CATCCTCAAT GTCATCAAAC GATCTGGTGT AATCGTTCGT CAGAAACTTC GCCCTGCTCC GTTACTCTCT GCAGACCATA AACTCAAGCG
     Tc3  CAAAAAGGAC CATCCTCAAT GTCATCAAAC GATCTGGTGT AATCGTTCGT CAGAAACTTC GCCCTGCTCC GTTACTCTCT GCAGACCATA AACTCAAGCG
T25G12RC  CTAAACAAAC AGTCCTGAAT ATGCTCAGTC AAAATCCTTC CATCGTGCGA CAGAAAATGA AAAAGGCTCC ATCAATGACG CCCGATCACA TGCTCAAACG
    ZC64  CTAAACAAAC AGTCCTGAAT ATGCTCAGTC AAAATCCTTC CATCGTGCGA CAGAAAATGA AAAAGGCTCC ATCAATGACG CCCGATCACA TGCTCAAACG
   C25G4  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........

          1101                                                                                                                  1200
   F27E5  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  T13A10  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
   K10B3  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
ZK1086RC  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
 R10H1RC  ATTGGAATTT GCTAAGAACA ATATGGGAAC GAATTGGAGT AAAGTGAGAA TTTAAAAAAG CAAGAGTGAA TAATTAGGAT CATTGTTTTA GGTTGTCTTC
   T02G5  ATTGGAATTT GCTAAGAACA ATATGGGAAC GAATTGGAGT AAAGTGAGAA TTTAAAAAAG CAAGAGTGAA TAATTAGGAT CATTGTTTTA GGTTGTCTTC
   B0303  ATTGGAATTT GCTAAGAACA ATATGGGAAC GAATTGGAGT AAAGTGAGAA TTTAAAAAAG CAAGAGTGAA TAATTAGGAT CATTGTTTTA GGTTGTCTTC
     Tc3  ATTGGAATTT GCTAAGAACA ATATGGGAAC GAATTGGAGT AAAGTGAGAA TTTAAAAAAG CAAGAGTGAA TAATTAGGAT CATTGTTTTA GGTTGTCTTC
T25G12RC  TTTAGATTTT GCAAAGGAAA ATATGGGCAC CGCATGGACA AAGGTATGGA TCCAATTAAA AAAAAACAAC ATATT..... .TACTTTGCA GATTGTGTTT
    ZC64  TTTAGATTTT GCAAAGGAAA ATATGGGCAC CGCATGGACA AAGGTATGGA TCCAATTTAA AAAAAACAAC ATATT..... .TACTTTGCA GATTGTGTTT
   C25G4  .......... .......... .........C CGCTTTGCC. .......... .......... .......... .......... .......... ..........
```

```
          1201                                                                                                                  1300
   F27E5  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  T13A10  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
   K10B3  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
ZK1086RC  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
 R10H1RC  TCCGATGAAA AGAAATTCAA TCTCGATGGG CCTGACGGTT GCCGCTACTA TTGGCGCGAT TTGCGCAAGG AACCAATGGT TTTTTCGAGA CGTAATTTTG
   T02G5  TCCGATGAAA AGAAATTCAA TCTCGATGGG CCTGACGGTT GCCGCTACTA TTGGCGCGAT TTGCGCAAGG AACCAATGGT TTTTTCGAGA CGTAATTTTG
   B0303  TCCGATGAAA AGAAATTCAA TCTCGATGGG CCTGACGGTT GCCGCTACTA TTGGCGCGAT TTGCGCAAGG AACCAATGGT TTTTTCGAGA CGTAATTTTG
     Tc3  TCCGATGAAA AGAAATTCAA TCTCGATGGG CCTGACGGTT GCCGCTACTA TTGGCGCGAT TTGCGCAAGG AACCAATGGT TTTTTCGAGA CGTAATTTTG
T25G12RC  TCCGATGAGA AGAAATTCAA TTTGGATGGG CCGGACGGAA ACAGATTTTA CTGGAGAGAT TTGAGAAAAG ATCCAATGGT TTTTTCAAAA AGAAATTTCG
    ZC64  TCCGATGAGA AGAAATTCAA TTTGGATGGG CCGGACGGAA ACAGATTTTA CTGGAGAGAT TTGAGAAAAG ATCCAATGGT TTTTTCAAAA AGAAATTTCG
   C25G4  .......... .......... .......... .......... .......... .......... .......... ........GC TTTTTC.... ..........

          1301                                                                                                                  1400
   F27E5  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  T13A10  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
   K10B3  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
ZK1086RC  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
 R10H1RC  GAGGAGGAAC GGTGATGGTT TGGGGAGCGT TCACGGAGAA GAAGAAGCTT GAGATACAGT TCGTCAGTAG CAAGATGAAC AGCACTGACT ATCAGAACGT
   T02G5  GAGGAGGAAC GGTGATGGTT TGGGGAGCGT TCACGGAGAA GAAGAAGCTT GAGATACAGT TCGTCAGTAG CAAGATGAAC AGCACTGACT ATCAGAACGT
   B0303  GAGGAGGAAC GGTGATGGTT TGGGGAGCGT TCACGGAGAA GAAGAAGCTT GAGATACAGT TCGTCAGTAG CAAGATGAAC AGCACTGACT ATCAGAACGT
     Tc3  GAGGAGGAAC GGTGATGGTT TGGGGAGCGT TCACGGAGAA GAAGAAGCTT GAGATACAGT TGCTCAGTAG CAAGATGAAC AGCACTGACT ATCAGAACGT
T25G12RC  GTGGTGGATC GGTGATGGTT TGGGCCGCTT TTTCGGAAAA GAAGAAGCTC CCCATCCAAT TTACAAGTCA TAAAATGACT GCTGTAGATT ATCAGCAAGT
    ZC64  GTGGTGGATC GGTGATGGTT TGGGCCGCTT TTTCGGAAAA GAAGAAGCTC CCCATCCAAT TTACAAGTCA TAAAATGACT GCTGTAGATT ATCAGCAAGT
   C25G4  .......... ........GC TGGGCCGCTT TTTCGGAAAA GAAGAAGCTC CCCATCCAAT TTACAAGTCA TAAAATGACT GCTGTAGATT ATCAGCAAGT
```

```
          1401                                                                                                  1500
    F27E5 .......... .......... .......... .......... ...AAAAATT TACTTTTACC CTGAGTCTAT GAAATTATGC ATAATAACA. ..........
   T13A10 .......... .......... .......... .......... ...AAATATT TGGTTTTACC CTGAGTCTAT GAAATTATGC ATAATAACA. ..........
    K10B3 .......... ......TCCA AATATCTTCG TCACTACTCC AGAAAAGACT TTAGATTTCA GCAGGATAAT GCGACAATCC ATGTGAGCAA CTCAACCCGC
 ZK1086RC .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  R10H1RC CTTGGAACTG GAGCTCTCCA AATATCTTCG TCACTACTCC AGAAAAGACT TTAGATTTCA GCAGGATAAT GCGACAATCC ATGTGAGCAA CTCAACCCGC
    T02G5 CTTGGAACTG GAGCTCTCCA AATATCTTCG TCACTACTCC AGAAAAGACT TTAGATTTCA GCAGGATAAT GCGACAATCC ATGTGAGCAA CTCAACCCGC
    B0303 CTTGGAACTG GAGCTCTCCA AATATCTTCG TCACTACTCC AGAAAAGACT TTAGATTTCA GCAGGATAAT GCGACAATCC ATGTGAGCAA CTCAACCCGC
      Tc3 CTTGGAACTG GAGCTCTCCA AATATCTTCG TCACTACTCC AGAAAAAACT TTAGATTTCA GCAGGATAAT GCGACAATCC ATGTGAGCAA CTCAACCCGC
 T25G12RC CTTGGAAAAG GATTCAGTGA AGTTTTTGAG ACACCCGTCG AAAAAAAACT GGCAGTTCCA ACAGGACAAC GCCAGTATTC ATTCAGCCAA TTCAACTCGT
     ZC64 CTTGGAAAAG GATTTAGTGA AGTTTTTGAG ACACCCGTCG AAAAAAAACT GGCAGTTCCA ACAGGACAAC GCCAGTATTC ATTCAGCCAA TTCAACTCGT
    C25G4 CTTGGAAAAG GATTTAGTGA AGTTTTTGAG ACACCCGTCG AAAAAAAACT GGCAGTTCCA ACAGAACAAC GCCAGTATTC ATTCAGCCAA TTCAACTCGT

          1501                                                                                                  1600
    F27E5 .......TCA TTATTAGTAC AGAAAACTTG CTCTATTGAC AATTTTAT.. ......TCTG GTTCAACTCA GAGGCTGGTA AATCCTTATT GCTAATCACA
   T13A10 .......TCA CTTTTAGAAC GGAAAACTTG CCTTATTGAC AATTTCAT.. ......TCTG GGTCAGCTCA GAGGCTAGTA AATCCTTATT GCTGATGCCG
    K10B3 GACTATTTCA AGCTCAAGAA GATCAACCTT CTTGATTGGC CAGCTCGAAG TCCTGATCTC AATCCAATCG AAAATTTGTG GGGGATTCTT GTCCGTATCG
 ZK1086RC .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  R10H1RC GACTATTTCA AGCTCAAGAA GATCAACCTT CTTGATTGGC CAGCTCGAAG TCCTGATCTC AATCCAATCG AAAATTTGTG GGGGATTCTT GTCCGTATCG
    T02G5 GACTATTTCA AGCTCAAGAA GATCAACCTT CTTGATTGGC CAGCTCGAAG TCCTGATCTC AATCCAATCG AAAATTTGTG GGGGATTCTT GTCCGTATCG
    B0303 GACTATTTCA AGCTCAAGAA GATCAACCTT CTTGATTGGC CAGCTCGAAG TCCTGATCTC AATCCAATCG AAAATTTGTG GGGGATTCTT GTCCGTATCG
      Tc3 GACTATTTCA AGCTCAAGAA GATCAACCTT CTTGATTGGC CAGCTCGAAG TCCTGATCTC AATCCAATCG AAAATTTGTG GGGGATTCTT GTCCGTATCG
 T25G12RC GCCTTTCTCA GCAGCAAGAA AATTAAACTC CTTAAATGGC CAGCTTGTTC ACCGGACCTC AATCCGATCG AAAATATGTG GGCTTCCCTC GTGAGACTCG
     ZC64 GCCTTTCTCA GCAGCAAGAA AATTAAACTC CTTAAATAGC CAGCTTGTTC ACCGGACCTC AATCCGATCG AAAATATGTG GGCTTCCCTC GTGAGACTCG
    C25G4 GCCTTTCTCA GCAGCAAGAA AATTAAACTC CTTAAATGAC CAGCTTGTTC ACCGGACCTC AATCCGATCG AAAATATGTG GGCTTCCCTC GTGAGACTCG
```

```
         1601                                                                                                                  1700
   F27E5 GATACACTCA .......... ....CAATGA TTGTGAGAGT TCATTGAACA ATTTTCAATG TCGGGGGTTC ...ACCGGGA CCCGAGTTAT ACACACTG..
  T13A10 GATACGGTCA .......... ....CAATGA TTGTGAGAGT TTATTAAACA ATTTTCAAGA TTAGGGGTTC ...ACCGGGA CCCGATTAGT ACTCACTG..
   K10B3 TGTATGCTCA GAACAAGACT TACCCAACAG TTGCATCGTT GAAGCAAGGA ATTCTCGACG CTTGGAAGTC TATTCCGGAC AACCAGCTGA AAAGTTTGGT
ZK1086RC .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
 R10H1RC TGTATGCTCA GAACAAGACT TACCCAACAG TTGCATCGTT GAAGCAAGGA ATTCTCGACG CTTGGAAGTC TATTCCGGAC AACCAGCTGA AAAGTTTGGT
   T02G5 TGTATGCTCA GAACAAGACT TACCCAACAG TTGCATCGTT GAAGCAAGGA ATTCTCGACG CTTGGAAGTC TATTCCGGAC AACCAGCTGA AAAGTTTGGT
   B0303 TGTATGCTCA GAACAAGACT TACCCAACAG TTGCATCGTT GAAGCAAGGA ATTCTCGACG CTTGGAAGTC TATTCCGGAC AACCAGCTGA AAAGTTTGGT
     Tc3 TGTATGCTCA GAACAAGACT TACCCAACAG TTGCATCGTT GAAGCAAGGA ATTCTCGACG CTTGGAAGTC TATTCCGGAC AACCAGCTGA AAAGTTTGGT
T25G12RC TGTACGCTAA TGGAAAACAG TATCCGAATG TTGCTGCTCT TAAAGTCGGA ATTGAGGATT CATGGAACGC CATATCAGCT ACAGAGATGA AAAATCTGGT
    ZC64 TGTACGCTAA TGGAAAACAG TATCCGAATG TTGCTGCTCT TAAAGTCGGA ATTGAGGATT CATGGAACGC CATATCAGCT ACAGAGAT.. ..........
   C25G4 TGTACGCTAA TGGAAAACAA TATCCGAATG TTGCTGCTCT TAAAGTCGGA ATTGAGGATT CATGGAACGC CATATCAGCT ACAGAGATGA AAAATCTGGT

         1701                                                                                                                  1800
   F27E5 .......... .......... .......... .......AAA CGGAGAAACG GCCTGAAAAA TGAGGCCCAT GTACGGTT.. .....TCAGC GGTGCAGCGG
  T13A10 .......... .......... .......... .......AAA CGGAGAAATG GCCTGAAATA ATAGGCCCAT ....GGTT.. .....TCAGC GGTGCAGCGG
   K10B3 CAGATCAATG GAGGACAGAC TGTTTGAGAT CATCCGCACA CAAGGAAACC CGATTAACTA TTGATCCTTT CTTGATTTTA GTATATGAAT GTTCTGTTGT
ZK1086RC .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
 R10H1RC CAGATCAATG GAGGACAGAC TGTTTGAGAT CATCCGCACA CAAGGAAACC CGATTAACTA TTGATCCTTT CTTGATTTTA GTATATGAAT GTTCTGTTGT
   T02G5 CAGATCAATG GAGGACAGAC TGTTTGAGAT CATCCGCACA CAAGGAAACC CGATTAACTA TTGATCCTTT CTTGATTTTA GTATATGAAT GTTCTGTTGT
   B0303 CAGATCAATG GAGGACAGAC TGATTGAGAT CATCCGCACA CAAGGAAACC CGATTAACTA TTGATCCTTT CTTGATTTTA GTATATGAAT GTTCTGTTGT
     Tc3 CAGATCAATG GAGGACAGAC TGTTTGAGAT CATCCGCACA CAAGGAAACC CGATTAACTA TTGATCCTTT CTTGATTTTA GTATATGAAT GTTCTGTTGT
T25G12RC CAATTCGATG CCTAATCGAA TCTTTGAGGT CATCGCCAAG AATGGAGGTC CTACGAAATA TTTAACTTTA TCTAAGTTAA TAAAATCTGT TGTGTTTTTT
    ZC64 .......... .......... .......... ....GCCAAG AATGGAGGTC CTACGAAATA TTGAACTTTA TCTAAGTTAA TAAAATCTGT TGTGTTTTTT
   C25G4 CAATTCGATG CCTAATCGAA TCTTTGAGGT CATCGCCAAG AATGGAGGTC CTACGAAATA TTGAACTTTA TCTAAGTTAA TAAAATCTGT TGTGTTTTTT
```

```
          1801                                                                                                                  1900
   F27E5  GGCTCAAACT TTTCGGATTT TTCCAAAAA. .......... ........CT TAGGGATGTT GTTAGAACAG TGTCATGTAT CATTTTGTAA AGTTTGGTAG
  T13A10  GGCTCAAACT TTTCGGATTT TTCCAAAAA. .......... ........CG TACGAATGTT GTTAGAACAG TGTCATGTCT CATTTTATAA AGTGTGGTTG
   K10B3  TGATCAAAAA TAACTGCAAC TTGTTAATAC GCTGTTTCTG ACTGGTTTCT TGGGGATGGC GTAAAAATGT TTATGGTGTG TGTGCTAGGA ATTTTAGTAG
ZK1086RC  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
 R10H1RC  TGATCAAAAA TAACTGCAAC TTGTTAATAC GCTGTTTCTG ACTGGTTTCT TGGGGATGGC GTAAAAATGT TTATGGTGTG TGTGCTAGGA ATTTTAGTAG
   T02G5  TGATCAAAAA TAACTGCAAC TTGTTAATAC GCTGTTTCTG ACTGGTTTCT TGGGGATGGC GTAAAAATGT TTATGGTGTG TGTGCTAGGA ATTTTAGTAG
   B0303  TGATCAAAAA TAACTGCAAC TTGTTAATAC GCTGTTTCTG ACTGGTTTCT TGGGGATGGC GTAAAAATGT TTATGGTGTG TGTGCTAGGA ATTTTAGTAG
     Tc3  TGATCAAAAA TAACTGCAAC TTGTTAATAC GCTGTTTCTG ACTGGTTTCT TGGGGATGGC GTAAAAATGT TTATGGTGTG TGTGCTAGGA ATTTTAGTAG
T25G12RC  GATTTCTAGA GGATGGTGAA TGCGCGAAGG CCAGTAGTGC TATCTCGTAC TAGTAGTATA ATAAAAATAA GAGTTGCAAA CCTTTAAAAA ATTTTCAGAC
    ZC64  GATTTCTAGA GGATGGTGAA TGCGCGAAGG CCAGTAGTGC TATCTCGTAC TAGTAGTATA ATAAAAATAA GAGTTGCAAA CCTTTAAAAA AATTTCAGAC
   C25G4  GATTTCTAGA GGATGGTGAA TGCGCGAAGG CCAGTAGTGC TATCTCGTAC TAGTAGTATA ATAAAAATAA GAGTTGCAAA TCTTTAAAAA AATTGCAGAC

          1901                                                                                                                  2000
   F27E5  TGATAGGGCA GAT....... .......... .......... .......... ........GA CACGTGGAAA CTTCTGTTAC AG........ GACTGTTAGA
  T13A10  TGATAGGGTA AGT....... .......... .......... .......... ........GA CACGTGGAAA CTTCTGTTAC AT........ GACTGTTACA
   K10B3  GAACTTGTTA AAGAGAGTTT ATCAAGTTTT GGCATTACCG AG........ AGCACCAGGA AGCTTGAAAC AGACTTGAAC ATCAAAGATC GGAAGAATCA
ZK1086RC  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
 R10H1RC  GAACTTGTTA AAGAGAGTTT ATCAAGTTTT GGCATTACCG AG........ AGCACCAGGA AGCTTGAAAC AGACTTGAAC ATCAAAGATC GGAAGAATCA
   T02G5  GAACTTGTTA AAGAGAGTTT ATCAAGTTTT GGCATTACCG AG........ AGCACCAGGA AGCTTGAAAC AGACTTGAAC ATCAAAGATC GGAAGAATCA
   B0303  GAACTTGTTA AAGAGAGTTT ATCAAGTTTT GGCATTACCG AG........ AGCACCAGGA AGCTTGAAAC AGACTTGAAC ATCAAAGATC GGAAGAATCA
     Tc3  GAACTTGTTA AAGAGAGTTT ATCAAGTTTT GGCATTACCG AG........ AGCACCAGGA AGCTTGAAAC AGACTTGAAC ATCAAAGATC GGAAGAATCA
T25G12RC  CGATGAAATA GATGTTGTAG ATCATCTTAG TTGATTCTGG TGCTGATTCA AGTATTGGGC GCCGCGCAGC GCCCTCTAAC AT..AAAATT AGGCGTTTCA
    ZC64  CGATGAAATA GATGTTGTAG ATCATCTTAG TTGATTCTGG TGCTGATTCA AGTATTGGGC GC..CGCAGC GCCCTCTAAC AT..AAAATT AGGCGTTTCA
   C25G4  CGATGAAATA GATGTTGTAG ATCATCTTAG TTGATTCTGG TGCTGATTCA AGTATTGGGC GCCGCGCAAC GCCCTCTAAC AT..AAAATT AGGCGTTTCA
```

```
          2001                                                                                                    2100
   F27E5  GGACAGTTGT GGGTCAACCA CCA....... .......... AGTGCCAATA TGAAATTGCT TGAAA..... .....CTGTG TTTATTCTCA TTTTCATAAC
  T13A10  GGACAGTTGC GGGTCAACCC TCA....... .......... AATGCCAATA GGAAATTGCT TGAAA..... .....CTGTG TTTATTCTCA TTTTCATAAT
   K10B3  GAACCAGTCT GGAGATCCCA CGAGAACATT TCGTATTAAT TTTTTAAAGT CAAAACTGCT TATCTGAA.. .....CTCAT TTCTAACTCA A..ACTTGAT
ZK1086RC  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
 R10H1RC  GAACCAGTCT GGAGATCCCA CGAGAACATT TCGTATTAAT TTTTTAAAGT CAAAACTGCT TATCTGAA.. .....CTCAT TTCTAACTCA A..ACTTGAT
   T02G5  GAACCAGTCT GGAGATCCCA CGAGAACATT TCGTATTAAT TTTTTAAAGT CAAAACTGCT TATCTGAA.. .....CTCAT TTCTAACTCA A..ACTTGAT
   B0303  GAACCAGTCT GGAGATCCCA CGAGAACATT TCGTATTAAT TTTTTAAAGT CAAAACTGCT TATCTGAA.. .....CTCAT TTCTAACTCA A..ACTTGAT
     Tc3  GAACCAGTCT GGAGATCCCA CGAGAACATT TCGTATTAA. TTTTTAAAGT CAAAACTGCT TATATGAA.. .....CTCAT TTCTAACTCA A..ACTTGAT
T25G12RC  GGACAATTTT AGAGATCCTA CGA....ATT TCCACGGATA AATTTCTTTT GAAAATTGAT TTGCAAGATG TCTGGATCAG TTTGAACTTA AGTCATGAAT
    ZC64  GGACAATTTT AGAGATCCTA CGA....ATT TCCACGGATA AATTTCTTTT GAAAATTGAT TTGCAAGATG TCTGGATCAG TTTGAACTTA AGTCATGAAT
   C25G4  GGACAATTTT AGAGATCCTA TGA....TTT TCCACGGATA AATTTCTTCT GAAAATTGAT TTGCAAGATG TCTGGATCAG TTTGAACTTA AGTCATGAAT

          2101                                                                                                    2200
   F27E5  TATTTACGTG ATATTAACTT CAAAAACTAT TTTTTTCCGG TGAAATCGTC AAAAATA... .........T GAGGGGGTCC TATAGAAGTT GCACACCTGA
  T13A10  CATTTACGGG GTATTCATTC CAAAAACGCT TTTTTCTCAA TGAAATCGTC AAAAATA... .........T AAGGGGGTCC TATAGAAGTT GCACACCTTA
   K10B3  TTGAGCTTTG TCAAAAATTC CAAAATCTGT TGATTTTAGT AAAAATCGTG AAAAAAAACG GGGGGGGGGG GGGGGGGTCC TATAGAAGTT TCACACTGGA
ZK1086RC  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
 R10H1RC  TTGAGCTTTG TCAAAAATTC CAAAATCTGT TGATTTTAGT AAAAATCGTG AAAAAAAAC. .......... .GGGGGGTCC TATAGAAGTT TCACACTGGA
   T02G5  TTGAGCTTTG TCAAAAATTC CAAAATCTGT TGATTTTAGT AAAAATCGTG AAAAAAAAC. .......... .GGGGGGTCC TATAGAAGTT TCACACTGGA
   B0303  TTGAGCTTTG TCAAAAATTC CAAAATCTGT TGATTTTAGT AAAAATCGTG AAAAAAAAC. .......... .GGGGGGTCC TATAGAAGTT TCACACTGGA
     Tc3  TTGAGCTTTG TCAAAAATTC CAAAATCTGT TGATTTTAGT AAAAATCGTG AAAAAAAAC. .......... .GGGGGGTCC TATAGAAGTT TCACACTGGA
T25G12RC  AAGTTGTTCT TAGAAAAAGC CAAAAAAAGG GGATTTTATT AG.AAAAATC AAAAA..... .......TTC GAGGGGGCGC TATAGAAGTT GGGCGGTCGT
    ZC64  AAGTTGTTCT TAGAAAAAGC CAAAAAAAGG GGATTTTATT AG.AAAAATC AAAAA..... .......TTC GAGGGGGCGC TATAGAAGTT GGGCGGTCGT
   C25G4  AAGTTGTTCT TAGAAAAAGC CAAAAAAAGG GGATTTTATT AGAAAAAATC AAAAA..... .......TTC GAGGGGGCGC TATAGAAGTT GGGCGGTCGT
```

```
           2201                                                                                                     2300
    F27E5  TTTTCCAACT TTTTTCCGTA CGCAGC.CAG TTTTAGAGGT ACAAGTCTGA AAATTTTTGT GTCCACATAT TTTGACATTT TGAATCAATC TACGGAAAAA
   T13A10  TTTTCCAACA TTTTTTGGTA CGCAGC.CAG TTTTAGAGGT ACACCTCTGA AAATTTTTGT GCCCACATAT TTTGACATTT TGAATCAATC CACGGTAAAA
    K10B3  TTTTGA.CGT TTTTTCCAGC GCGACCTCAG TTTTTGAGAT AAACTTCTGG AAAAAAATGT ATATAGGTTT TTTTCAATTT AAAAACGAAT GGCATTAGCA
 ZK1086RC  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  R10H1RC  TTTTGA.CGT TTTTTCCAGC GCGACCTCAG TTTTTGAGAT AAACTTCTGG AAAAAAATGT ATATAGGTTT TTTTCAATTT AAAAACGAAT GGCATTAGCA
    T02G5  TTTTGA.CGT TTTTTCCAGC GCGACCTCAG TTTTTGAGAT AAACTTCTGG AAAAAAATGT ATATAGGTTT TTTTCAATTT AAAAACGAAT GGCATTAGCA
    B0303  TTTTGA.CGT TTTTTCCAGC GCGACCTCAG TTTTTGAGAT AAACTTCTGG AAAAAAATGT ATATAGGTTT TTTTCAATTT AAAAACGAAT GGCATTAGCA
      Tc3  TTTTGA.CGT TTTTTCCAGC GCGACCTCAG TTTTTGAGAT AAACTTCTGG AAAAAAATGT ATATAGGTTT TTTTCAATTT AAAAACGAAT GGCATTAGCA
 T25G12RC  TTTAGAGACT TTTTGACCTA GTCTCCTCAG TCTATTGAAT TTTTTGATGA AAATTTTGGA CAGGGGGTTT TTCGACATTT GGAATCAAAT GGCATTAGCA
     ZC64  TTTAGAGACT TTTTGACCTA GTCTCCTCAG TCTATTGAAT TTTTTGATGA AAATTTTGGA CAGGGGGTTT TTCGACATTT GGAATCAAAT GGCATTAGCA
    C25G4  TTTAGAGACT TTTTGACCTA GTCTCCTCAG TCTATTGAAT TTTTTGATGA AAATTTTGGA CAGGGGGTTT TTCGACATTT GGAATCGAAT GGCATTAGCA

           2301                                                                           2376
    F27E5  ATTTTTCAAA AAAAATTTTC AATATAGGTG AAACGTCAAA AATGAGGGGG GTCCTATAGA ACTTTCTCAC ACTG..
   T13A10  TTTTTCAAAA AAAAATTTTC AAAATAGGTG AAGCGTCAAA AATGAAGGGG GTCCTATAGA ACTTTCTCAC ACTG..
    K10B3  GTTTTTCAAA AAAATTTTCG GAAGTTCCTC AAACCTTCAA ATTAGGGGGG GTCCTATAGA ACTTTCCCAC ACTG..
 ZK1086RC  .......... .......... .......... .......... .......... .......... .......... ......
  R10H1RC  GTTTTTCAAA AAAATTTTCG GAAGTTCCTC AAACCTTCAA ATTAGGGGGG GTCCTATAGA ACTTTCCCAC ACTG..
    T02G5  GTTTTTCAAA AAAATTTTCG GAAGTTCCTC AAACCTTCAA ATTAGGGGGG GTCCTATAGA ACTTTCCCAC ACTG..
    B0303  GTTTTTCAAA AAAATTTTCG GAAGTTCCTC AAACCTTCAA ATTAGGGGGG GTCCTATAGA ACTTTCCCAC ACTG..
      Tc3  GTTTTTCAAA AAAATTTTCG GAAGTTCCTC AAACCTTCAA ATTAGGGGGG GTCCTATAGA ACTTTCCCAC ACTG..
 T25G12RC  GTTTTTTGTA AAAAATGTCA GAAGGTGGTG AAATGAGAAA AAATAGGGGG GCGCTATAGA AGTTGGGCGC ACTG..
     ZC64  GTTTTTTGTA AAAAATGTCA GAAGGTGGTG AAATGAGAAA AAATAGGGGG GCGCTATAGA AGTTGGGCGC ACTG..
    C25G4  G.TTTTTGTA AAAAATGTCA GAAGGTGGTG AAATGAGAAA AAATA.GGGG GCGCTATAGA AGTTGGGCGC ACTGTA
```

APPENDIX E: Alignment of Tc4 and six cosmid sequences identified as high scoring blast hits to Tc4.

```
          1                                                                                                                  100
   F23c11 ATGATGTTAC TTACTAGGGA ATGACCAGAA TAAGTGGAGC GATATTCAAA AAAAAAATAT TGTATCGGAA AGCTGACATT CTCTACTATA AGAATATGAC
    Zk686 CTCTTCTGCC TCACTAGGGA ATGACCAGAA TAAGTGGAGC GATATTCAAA AAAAAAATAT TGTATCGGAA AGCTGACATT CTCTACTATA AGAATATGAC
 F49e11rc ..TATT.GAC TAACTAGGGA GTGTTTTAAC TATACGGTGC GATCGGGTAA AAGTAAACGT GTTATGCGAT AGCTGGCATC TTAGGCTTTC AGAAT....C
   F57g12 AGTATTACTC TTACTAGGGA GTGTTTTAAC TATGTGGTGC GATCGGGTAA AAGTAAACGT GTTATGCGAT AGCTGGCGTT TTAGGCTTTC AGAAT....C
  R04b3rc ...AAATAAC TGACTAGGGA GTGTTTTAAC TATACGGTGC GATCGAGTAA AAGTAAACGT GTCATGCGAA AGCTGGCAAT TTAGGCTTTC AGTAT....C
    T08g2 GCCTTTTGGC TGACTTGGGA ATGACCAGAA TAAGTGGAGC GATATTGAAA AAAAAA..AT TGTATCGGAA AGCTGACATT CTCTACTATA AGAATATGAC
      Tc4 .......... TAACTAGGGA ATGACCAGAA TAAATGGAGC GATATTCAAA AAAAAAATAT TGTATCGGAA AGCTGACATT CTCTACTATA AGAATATGAC

          101                                                                                                                200
   F23c11 TGAAATTTTT GCCCATTCGG G....CTGGA AATCTGAAAT TTTTACGTCT GAAATTCTAC .......... .......... ....GTTAAC .......TCT
    Zk686 TGAAATTTTT GCCCATTCGG G....CTGGA AATCTGAAAT TTTTACGTCT GAAATTCTAT .......... .......... ....GTTAAC .......TCT
 F49e11rc TGTAATTTGT TTCGGCAGAA GACCTCTGTG AGTCTGGAAA TTTTCATCTG AAAATGTAGT ACTGAAATCA GTGCATTTCC TATGGTTAAC AGTGGA.TTT
   F57g12 TGTAATTTGT TTCGGCAGAA GACCTCTGTG AGTCTGGAAA TTTTCATCTG AAAATGTAGT ACTGAAATCA GTGCATTTCC TATGGTTAAC AGTGGA.TTT
  R04b3rc TGTAATTTGT TCCGGCGGAA GACCTCTGTG AGTCTGGAAA TTTTCATCTG AAAATTTAGT ACTGAAATCA GTGCATTTCC TGTGGTTAAC AGTGGATTTT
    T08g2 TGAAATTTTT GCCCATTCGG G......... .......... .......... .......... .......... .......... .......... ..........
      Tc4 TGAAATTTTT GCCCGTTCGG G....CTGGA AATCTGAAAT TTTTACGTCT GAAATTCTAC ACTGAAATCA GTGCATTTCC TATGGTTAAC AGTGGATTTT

          201                                                                                                                300
   F23c11 TTTCACTATC CCCA..ATTA GTACTGCC.. .....TGCAA CAGCGA.... ........GA TGGCCGAGTG GATAGAG... .......TGG ATGACAAGCA
    Zk686 TTTCACTATC CCCA..ATTA GTACTGCC.. .....TGCAA CAGCGA.... ........GA TGGCCGAGTG GATAGAG... .......TGG ATGACAAGCA
 F49e11rc TGTCTCTGGC GCCAACAGAA GTCTCACCAC AATGGTGGAA GGGCGAAATC ATCGCTTCGG TGGTCGAGTG GTGAACGCGT TCGCCTCTTG AGCAGAAGTT
   F57g12 TGTCTCTGGC GCCAACAGAA GTCTCACCAC AATGGTGGAA GGGTGAAATC ATCGCTTCGG TGGTCGAGTG GTGAACGCAT TCGCCTCTTG AGCAAAAGTT
  R04b3rc TGTCTCTGGC GCCAACAGAA GTCTCACCAC AATGGTGGAA GGGCGAAAAC ATCGCTTCGG TGGTCGAGTG GTGAACGCGT TCGCCTCTTG AGCAGAAGTT
    T08g2 .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
      Tc4 TGTCTCTGGC GCCAACAGAA GTCTCACCAC AATGGT.GAA GGGCGAAAAC ATCGGTTCGG TGGTCGAGTG GTGAACGCGT TCGCCTCTTG AGCAGAAGTT
```

```
          301                                                                                                       400
  F23c11  TGGTGG.... .....GACTC TGGGGTTCAA TTCTACCC.. ...TAACGTA AAATTTTT.. .......... .......... .......... ..........
   Zk686  TGGTGG.... .....GACTC TGGGGTTCAA TTCCACCC.. ...TAACGTA AAATTTTT.. .......... .......... .......... ..........
F49e11rc  TGTGGGTTCG GTTCCCACAC ATGGTTTAAA TTTTGGCC.T TTTTTATACA AAATTTTTAG AACGGGAAAC .......... .......... ..........
  F57g12  TGTGGGTTCG GTTCCCACAC ATGGTTTAAA TTTTGGCCTT TTTTTATACA AAATTTTTAG AACGGGAAAC .......... .......... ..........
 R04b3rc  TGTGGGTTCG GTTCCCACAC ATGGTTTAAA TTTTGGCCTT TTTTTATACA AAATTTTTAG AACGGGAAAC .......... .......... ..........
   T08g2  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
     Tc4  TGTGGGTTCG GTTCCCATAC ATGGTTTAAC TTTTGGCC.T TTTTTATACA AAATTTTCAG AACGGGAAAC AAGTATTTAG AACATTTTTT TGAGGGTTTT

          401                                                                                                       500
  F23c11  .......... .......... .......... .......... .CAAAAATTT AAACGTGTTT .......... .......... .......... ..........
   Zk686  .......... .......... .......... .......... .CAAAAATTT AAACGTGTTT .......... .......... .......... ..........
F49e11rc  .......... .......... .......... .......... ..AAATGTTT AAAACAGTTT .......... .......... .......... ..........
  F57g12  .......... .......... .......... .......... ..AAATGTTT AAAACTGTTT .......... .......... .......... ..........
 R04b3rc  .......... .......... .......... .......... ..AAATGTTT AAAACAGTTT .......... .......... .......... ..........
   T08g2  .......... .......... .......... ........CT GGAAACTTTT CAGAAATTTT .......... .......... .......... ..........
     Tc4  ACATAATTTT TTTGCTTTTT AATTGAACCA TAATTACCCT GGAAACTTTT CAGAAATTTT AATTTTTTTC GAAAATTGTC ACTTTTTTCT CCACCAAACC

          501                                                                                                       600
  F23c11  .......... .......... ........TT TCGAAACTAT ATAAA..... .......... .......... .......AGC CCAAATTTAA ..........
   Zk686  .......... .......... ........TT TCGAAACTAT ATAAA..... .......... .......... .......AGC CCAAATTTAA ..........
F49e11rc  .......... .......... .........T TTGAGGTTTT .......... .......... .......... ......TACA TTACTTTTTT GCTTTTTGAT
  F57g12  .......... .......... .........T TTGAGGTTTT .......... .......... .......... ......TACA TTACTTTTTT GCTTTTTGAT
 R04b3rc  .......... .......... .........T TTGAAATTTT TTAAAAAATG CATGAAATAT TTTAGAGTGT CACAAATAAC CTATTTTTCA TTATTTTCAA
   T08g2  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
     Tc4  CATGAGAAAA TTTGATCGAA AAATTTTTTT TTGAAATTTT TTAAAAAATG CATGAAATAT TTTAGAGTGT CACAAATAAC CTATTTTTCA TTATTTTCAA

          601                                                                                                       700
  F23c11  .......... .......... .......... .......TTT TTCTGATCGA TATCAGCATG A......... ........TC AATGTGTC.. ..........
   Zk686  .......... .......... .......... .......TTT TTCTGATCGA TATCAGCATG A......... ........TC AATGTGTC.. ..........
F49e11rc  TGAACCA..C AATTACCCTG GAAACTTTTC AGAAATTTTT ATTTTTTCGA AAATTGCCA. ....CTTTTT T......CTC C......... ..........
  F57g12  TGAACCA..C AATTACCCTG GAAACTTTTC AGAAATTTTT ATTTTTTCGA AAATTGCCA. ....CTTTTT T......CTC CACCAAACCC ATGAGAAAAT
 R04b3rc  TGACCGAATC ACTGATTCTG ATGCCTTATC AAGACGTTTT ACCAAATCGA TATTGGCAAA ACATCTTGTT TTTGAGACTC CATATCTCCG CAGGAAAAAT
   T08g2  .......... .......... .......... ......TATT ACCAAATCGA TATTGGAAAA ACACCTTGTC TTTGAGACTC CATATCTCTG CAGGAAAAAA
     Tc4  TGACCGAATC ATTGATTCTG ATGCCTTATC AAGACGTTTT ACCAAATCGA TATTGGCAAA ACATCTTGTT TTTGAGGCTC CATATCTCTG CAGGAAAAAA
```

```
          701                                                                                                       800
  F23c11  .......... .......... .......... .......GAC CAACACCAT. .........A ATGATTCATT CAACATCTTT TCCCCTAAAA T.........
   Zk686  .......... .......... .......... .......GAC CAACACCAT. .........A ATGATTCATT CAACATCTTT TCCCCTAAAA T.........
F49e11rc  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  F57g12  TGGTTCCAAA AATTTTTTTT GAAATTTTTT CAAAAATGTA TGAAATATTG TAGAGTGTCA CAAATAACTT AAACACTTTC TTGTAAAAAA TTTGGTTGCC
 R04b3rc  TGCGCTTAAA AGT....... .GATCAACTG AAAACTTGTT AAACACAATG TGATCTAAAA CTTTTCAGTT GAACACTTTT TTGTAAAAAA TTTGGTTGCC
   T08g2  TCGCACTAAA AAGT...... .GATCAACTG AAAACTTGTT AAACACAATG TGATCTAAAA CTTTTTAGTT GAACACTTTC TTGTAAAAAA CTTGGTTGCC
     Tc4  TCGCACTAAA AAGT...... .GATTAACTG AAAACTTGTT AAACACAATG TAATCTAAAA CTTTTCAGTT GAACACTTTT TTGTAAAAAA TTTCGTTGCC

          801                                                                                                       900
  F23c11  .......... ..TTAAACTT TTT......T TCTCAAA.AT TGTAACTACA ACCA..TGGA TCC.CCTCAT ATAAACAATT TTATTCA... ........TA
   Zk686  .......... ..TTAAACTT TTT......T TCTCAAA.AT TGTAACTACA ACCA..TGGA TCC.CCTCAT ATAAACAATT TTATTCA... ........TA
F49e11rc  .AGATATAGA CCTTTAACTA TTTAGAATAT TCAAAAATAA TGAAGCTCAA ATCAATTGGT TCCAACTCGG CAACCAAATT TTTTACAAAA AAGTGTTCAA
  F57g12  AAGATATAGA CCTTTAACTA TTTAGAAAAT TCAAAAATAA TGAAGCTCAA ATCAATTGGT TCCAACTCGG CAACCAAATT TTTTACAAGA AAGTGTTTTA
 R04b3rc  AAGATATGGA CCGTTAACTA TTTAGAATAT TCAAAAATAA TGAAGCTCAA ATCAATTGGT TCCAACTCGG CAACCAAAAT TTTTACAAGA AAGTGTTCAA
   T08g2  AAGATATAGA CCTTTAACTA TTTAGAATAT TCAAAAATAA TGAAGCTCGA ATCAATTGGT CCCAACTCGG CAACCAAATT TTTTTAAAGA AAGTTTTTTA
     Tc4  AAGATATAGA TCTTTAACTA TTTAGAATAT T......... ..GAGCTCAA ATCAATTGGT TCCAACTCGG CAACGAAATT TTTTACAAAA AAGTGTTCAA

          901                                                                                                      1000
  F23c11  GCGAATAATA TGAAAAAAAA A..TTTAAAA AAAGTTTAAA .......... TTTTAGGG.. .......... GAAAAGATGT TGA....... ........AT
   Zk686  GCGAATAATA TGAAAAAAAA T..TTTTAAA AAAGTTTAAA .......... TTTTAGGG.. .......... GAAAAGATGT TGA....... ........AT
F49e11rc  CTGAAAAGTT TTAGATCACA TTGTGTTTAA CAAGTTTTCA GTTGATCACT TTTAAGCGCA ATTTTTTCCT GCGGAGATAT GGAGTCTC.A AAGACAAGGT
  F57g12  CTGAAAAGTT TTAGATCACA TTATTTTCAA CAAGTTTCTA GTTGATCACT TTTTAGTGCG ATTTTTTTCT GCGGAGATAT GGAGTCTT.A AAGACAAGGT
 R04b3rc  CTGAAAAGTT TTAGATCACA TTGTGTTTAA CAAGGTTATA GTTGATCACT TTTTAGTGCG ATTTTTTCCT GCAGAGATAT GGAGCCTC.A AAAACAAGAT
   T08g2  CTGAAAAGTT TCAGATCACA TTATGTTTAA CAAGTTTCTA GTTGATCACT TTTTAGTGCA ATTTTTTCCT GCAGAGATAT GGAGTCTC.A AAGACAAGGT
     Tc4  CTGAAATGTT TTAGATCACA TTGTGTTTAA CAAGTTTTCA GTTAATCACT TTTTAGTGCG ATTTTTT.CT GCAGAGATAT GGAGCCTCAA AAAACAAGAT

          1001                                                                                                     1100
  F23c11  GAAT...CAT TATGGTGTTG GT.CGACACA TTGATCATGC .......... .......... ...TGATATC GACCAGAAAA ATTAAATTTG GG........
   Zk686  GAAT...CAT TATGGTGTTG GT.CGACACA TTGATCATGC .......... .......... ...TGATATC GATCAGAAAA ATTAAATTTG GG........
F49e11rc  GTTTTGCCAA TATCGATTTG GTAAAACGTC TTGATACGGC ATCAGAATAA GTGATTCGGT CATTGAAAAC AATGAAAAAT AAGTTATTTG TGACACTCTA
  F57g12  GTTTTGCCAA TATCGATTTG GTAAAACGTC TTGATACGGC ATCAGAATAA GTGATTCGGT CATTGAAAAC AATGAAAAAT AAGTTATTTG TGACACTCTA
 R04b3rc  GTTTTGCCAA TATCGATTTG GTAAAACGTC TTGATAAGGC ATCAGAATCA GTGATTCGGT CATTGAAAAT AATGAAAAAT AGGTTATTTG TGACACTCTA
   T08g2  GTTTTTCCAA TATCGATTTG GTAAAACGTC TTGATACGGC ATCAGAATCA GTGATTCGGT CATTGAAAAC AATGAAAAAT AAGTTATTTG TGACACTCTA
     Tc4  GTTTTGCCAA TATCGATTTG GTAAAACGTC TTGATAAGGC ATCAGAATCA ATGATTCGGT CATTGAAAAT AATGAAAAAT AGGTTATTTG TGACACTCTA
```

```
          1101                                                                                                                  1200
F23c11    .......... .....CTTTT ATATAGTTTC GAAAAAACAC GTTT...... .......... .......... .......... .......... ..........
Zk686     .......... .....CTTTT ATATAGTTTC GAAAAAACAC GTTT...... .......... .......... .......... .......... ..........
F49e11rc  CAATATTTCA TACATTTTTG AAAAAATTTC .AAAAAAAAT TTTTGGAACC AATTTTCTCA TGGGTTTGGT GGAGAAAAAA GTGGCAATTT TCG.AAAAAA
F57g12    CAATATTTCA TACATTTTTG AAAAAATTTC .AAAAAAAAT TTTTGGAACC AATTTTCTCA TGGGTTTGGT GGAGAAAAAA GTGGCAATTT TCG.AAAAAA
R04b3rc   AAATATTTCA TGCATTTTTT AAAAAATTTC .AAAAAAAAT TTTTGAATCA AATTTTCTCA TGGGTTTGGT GGAGAAAAAA GTGACAATTT TCG.AAAAAA
T08g2     CATTATTTCG TACATTTTTG AAAAAATTTC .AAAAAAAAT TTTTGGATCA AATTTTCTCA TGGGTTTGGT GGAGAAAAAA GTGGCAATTT TCG.AAGAAA
Tc4       AAATATTTCA TGCATTTTTT AAAAAATTTC AAAAAAAAAT TTTTCGATCA AATTTTCTCA TGGGTTTGGT GGAGAAAAAA GTGACAATTT TCGAAAAAAA

          1201                                                                                                                  1300
F23c11    ...AAATTTT TGAAAAATTT TACG.....T TAGGGTGGAA TTGAA..... .......... .......... .......... .......... ..........
Zk686     ...AAATTTT TGAAAAATTT TACG.....T TAGGGTGGAA TTGAA..... .......... .......... .......... .......... ..........
F49e11rc  TAAAAATTTC TGAAAAGTTT CCAGGGTAAT TGTGGTTCAA TCAAAAAGCA AAAAAGTAAT GTAAAAACCT CAAAAAACTG TTTTAAACAT TTGTTTCCCG
F57g12    TAAAAATTTC TGAAAAGTTT CCAGGGTAAT TGTGGTTCAA TCAAAAAGCA AAAAAGTAAT GTAAAAACCT CTAAAAACTG TTTTAAACAT TTGTTTCCCG
R04b3rc   TTAAAATTTC TGAAAAGTTT CCAGTGTAAT TGTGGTTCAA TCAAAAAGCA AAAAAGTAAT GTAAAAACCT TAAAAAACTG TTTTAAACAT TTGTTTCCCG
T08g2     TAAAAATTTC TGAAAAG... .......... .......... .......... .......... .......... .......... .......... ..........
Tc4       TTAAAATTTC TGAAAAGTTT CCAGGGTAAT TATGGTTCAA TTAAAAAGCA AAAAAATTAT GTAAAACCCT CAAAAAAATG TTCTAAATAC TTGTTTCCCG

          1301                                                                                                                  1400
F23c11    .......... .......... .....CCCCA GAGTCCCACC ATGCTTGTCA TC........ .......... .......... ...CACTCTA TCCACTCGGC
Zk686     .......... .......... .....CCCCA GAGTCCCACC ATGCTTGTCA TC........ .......... .......... ...CACTCTA TCCACTCGGC
F49e11rc  TTCTAAAAAT TTTGTAT.AA AAAAGGCCAA AATTTAAACC ATGTGTGGGA ACCGAACCCA CAAACTTCTG CTCAAGAGGC GAACGCGTTC ACCACTCGAC
F57g12    TTCTAAAAGT TTTGTATAAA AAAAGGCCAA AATTTAAACC ATGTGTGGGA ACCGAACCCA CTAACTTCTG CTCAAGAGGC GAACGCGTTC ACCACTCGAC
R04b3rc   TTCTAAAAAT TTTGTATAAA AAAAGGCCAA ATTTTAAACC ATGTGTGGGA ACCGAACCCA CAAACTTCTG CTCAAGAGGC GAACGCGTTC ACCACTCGAC
T08g2     .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
Tc4       TTCTGAAAAT TTTGTAT.AA AAAAGGCCAA AAGTTAAACC ATGTATGGGA ACCGAACCCA CAAACTTCTG CTCAAGAGGC GAACGCGTTC ACCACTCGAC

          1401                                                                                                                  1500
F23c11    CA........ ....TCTCGC TGTTGCA... .......GGG AGTACTAATT GGGGATAGTG AAAAGA.... ...GTTAAC. .......... ..........
Zk686     CA........ ....TCTCGC TGTTGCA... .......GGG AGTACTAATT GGGGATAGTG AAAAGA.... ...GTTAAC. .......... ..........
F49e11rc  CACCGAAGCG ATGATTTCGC CCTTCCACCA TTGTGGTGAG ATTTCT.GTT GGCGCCAGAG AC.AAAATCC ACTGTTAACC ATAGGAAATG CACTGATTTC
F57g12    CACCGAAGCG ATGATTTCGC CCTTCCACCA TTGTGGTGAG ACTTCT.GTT GGCGCCAGAG AC.AAAATCC ACTGTTAACC ATAGGAAATG CACTGATTTC
R04b3rc   CACCGAAGCG ATGTTTTCGC CCTTCCACCA TTGTGGTGAG ACTTCT.GTT GGCGCCAGAG ACAAAAATCC ACTGTTAACC ATAGGAAATG CACTGATTTC
T08g2     .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
Tc4       CACCGAACCG ATGTTTTCGC CCTT.CACCA TTGTGGTGAG ACTTCT.GTT GGCGCCAGAG ACAAAAATCC ACTGTTAACC ATAGGAAATG CACTGATTTC
```

```
         1501                                                                                                                  1600
  F23c11 ....GTAGAA TTTCAGACGT AAAAATTTCA GATTTCCAGC CCGAATG... .........G GCAAAAATTT CAGTCATATT CTTATAGTAG AGAATGTCAG
   Zk686 ....GTAGAA TTTCAGACGT AAAAATTTCA GATTTCCAGC CCGAATG... .........G GCAAAAATTT CAGTCATATT CTTATAGTAG AGAATGTCAG
F49e11rc AGTACTACAT TTTCAGATGA AA........ .ATTTCCAGA CTCACAGAGG TCTTCTGCCG AAACAAATTA CAG....ATT CTGAAAGCCT AAGATGCCAG
  F57g12 AGTACTACAT TTTCAGATGA AA........ .ATTTCCAGA CTCACAGAGG TCTTCTGCCG AAACAAATTA CAG....ATT CTGAAAGCCT AAAACGCCAG
 R04b3rc AGTACTAAAT TTTCAGATGA AA........ .ATTTCCAGA CTCACAGAAG TCTTCCGCCG GAACAAATTA CAG....ATT CTGAAAGCCT AAGATGCCAG
   T08g2 .......... .......... .......... ..TTTCCAGC CCGAATG... .........G GCAAAAATTT CAGTCATATT CTTATAGTAG AGAATGTCAG
     Tc4 AGT.GTAGAA TTTCAGACGT AAAAATTTCA GATTTCCAGC CCGAACG... .........G GCAAAAATTT CAGTCATATT CTTATAGTAG AGAATGTCAG

         1601                                                                                                  1680
  F23c11 CTTTCCGATA CAAT.TTT.. .....TTTTT TTGAATATCG CTCCACTTAT TCTGGCCATT CCCTAGTTAG GTTTAA....
   Zk686 CTTTCCGATA CAATATTT.. .....TTTTT TTGAATATCG CTCCACTTAT TCTGGCCATT CCCTAGTCAG TCTCCTCCG.
F49e11rc CTATCGCATA ACACGTTTA. ......CTTT TACCCGATCG CACCGTATAG TTAAAACACT CCCTAGTAAT ACAATTTAC.
  F57g12 CTATCGCATA ACACGTTTAG GGAGTGTTTT AACTATGTGG TGCGATCGGG TAAAAGTAGT TAGTACTGTT GAACTCGAAA
 R04b3rc CTATCGCATG ACACGTTTA. ......CTTT TACCCGATCG CACCGTATAG TTAAAATACT CTCTAGTGAG TCATATCAA.
   T08g2 CTTTCCGATA CAATTTTT.. .....TTTTT TTCAATATCG CTCCACTTAT TCTGGTCATT CCCTAGTGAG CAGTAGTTA.
     Tc4 CTTTCCGATA CAATATTT.. .....TTTTT TTGAATATCG CTCCATTTAT TCTGGTCATT CCCTAG.... ..........
```

APPENDIX F: Alignment of Tc5 (upper) and the cosmid T13c2 (lower) sequence identified as high scoring blast hit to Tc5.

```
                 .         .         .         .         .
  1 ..........CAAGGGAAGGTTCTGAACTCGTTATCGGACTTCGTTACGC 40
             ||||||||||||||||||||||||||||||||||||||||
  1 CTGCCACTTACAAGGGAAGGTTCTGAACTCGTTATCGGACTTCGTTACGC 50
                 .         .         .         .         .
 41 CACTATATACATTCGATAGAGGATAGTTACAGATGATCCCTTCAAAAAAT 90
    ||||||||||||||||||||||||||||||||||||||| ||||||||
 51 CACTATATACATTCGATAGAGGATAGTTACAGATGATCCCTCCAAAAAAT 100
                 .         .         .         .         .
 91 TTAGCTGCTTCAGAGCAGGTTTGGCCAAGTTGTGACGTCTTGAATTTTGG 140
    |||||||||||||||||||||||||||||||||||||||| |||||
101 TTAGCTGCTTCAGAGCAGGTTTGGCCAAGTTGTGACGTCTTGAAGTTTGG 150
                 .         .         .         .         .
141 TGCTGAAATTCCTCATATCAAGTGATATTTCAATGACTACCACGCTGCAG 190
    ||||||||||||||||||||||||||||||||||||||||||||||||
151 TGCTGAAATTCCTCATATCAAGTGATATTTCAATGACTACCACGCTGCAG 200
                 .         .         .         .         .
191 AAACACCAGTGAACTCACCACTCTCAATTAGCGTTAGCAAACATGGCTTG 240
    ||||||||||||||||||||||||||||||||||||||||||||||||
201 AAACACCAGTGAACTCACCACTCTCAATTAGCGTTAGCAAACATGGCTTG 250
                 .         .         .         .         .
241 GTGGCCGAGTGGTAGTGGCGTGAGTTTCGAGGTGTGGTATTCGTGGTTCG 290
    ||||||||||||||||||||||||||||||||||||||||||||||||
251 GTGGCCGAGTGGTAGTGGCGTGAGTTTCGAGGTGTGGTATTCGTGGTTCG 300
                 .         .         .         .         .
291 GTTCCCCGTCAACATAAACTTTTTTTTTTAATTTTTAAAGTCAATCCATT 340
    ||||||||||||||||||||||||||||||||||||||||||||||||
301 GTTCCCCGTCAACATAAACTTTTTTTTTTAATTTTTAAAGTCAATCCATT 350
                 .         .         .         .         .
341 TCCAATTAGAACACATCTATAAACTTTTTCAAGTGGGAAAATGTGCAGAT 390
    |||||||||||||||||||||||||||||||||||||| |||||||
351 TCCAATTAGAACACATCTATAAACTTTTTCAAGTGGGAAAATATGCAGAT 400
                 .         .         .         .         .
```

```
391 ATTATCCCTATGAATCAAATGCGTCAATTCTCCAAATTTTTCCGA..... 435
    ||||||||||||||||||||||||||||||||||||||||||||
401 ATTATCCCTATGAATCAAATGCGTCAATTCTCCAAATTTTTCCGAAGTGT 450
                .         .         .         .         .
436 ..TTTTTTTTTTCAATATGTGTTAT........AGTTAAAAGCACAATAA 475
      |||||| ||  ||||  |||| |        | ||     |  |||
451 TTTTTTTTGTTGAAATAATTGTTTTTTTCACTGATTTTCTTCCGTAATTC 500
                .         .         .         .         .
476 AACAGATGTTTAAAGTA......CATACATTAAACATTAAATTTTCATTA 519
    || |  | ||||   ||       |   ||||||    |    | |||||
501 AAAATGTTTTTATTATATTTTATAAATGATTAAATGAAAGTAATACATTA 550
                .         .         .         .         .
520 AATTTTCAAATAATATCATCGTGGTTAAAAATGTAGGCCACAAGAAGAGC 569
    ||||||||||||||||||||||||||||||||||||||||||||||||||
551 AATTTTCAAATAATATCATCGTGGTTAAAAATGTAGGCCACAAGAAGAGC 600
                .         .         .         .         .
570 TGTTAGGTCCCACCACGCTTCACACCTTTCTTGTAGTTTTTTTT..TTGT 617
    ||||||||||||||||||||||||||||||||||||||||||  ||||
601 TGTTAGGTCCCACCACGCTTCACACCTTTCTTGTAGTTTTTTTTTGTTGT 650
                .         .         .         .         .
618 TATTTTCTGTTGACTCGTCTTCCGTTGTCTATATTTTAACTGAAAATGCC 667
    ||||||||||||||||||||||||||||||||||||||||||||||||||
651 TATTTTCTGTTGACTCGTCTTCCGTTGTCTATATTTTAACTGAAAATGCC 700
                .         .         .         .         .
668 CTTCCGCCCACAAGTAATCATCGGAGAACTTATGAAAACGTTTGGAACTA 717
    ||||||||||||||||||||||||||||||||||||||||||||||||||
701 CTTCCGCCCACAAGTAATCATCGGAGAACTTATGAAAACGTTTGGAACTA 750
                .         .         .         .         .
718 ATACAAACGCGTTGCCAATGAGTCGAGAAGAAACGAAAACTGTCGAGAAA 767
    ||||||||||||||||||||||||||||||||||||||||||||||||||
751 ATACAAACGCGTTGCCAATGAGTCGAGAAGAAACGAAAACTGTCGAGAAA 800
                .         .         .         .         .
768 TTTACAAGGATTCTCAAAGATGCTGAAACGGACGATCTTCTTATTCAAAG 817
    ||||||||||||||||||||||||||||||||||||||||||||||||||
801 TTTACAAGGATTCTCAAAGATGCTGAAACGGACGATCTTCTTATTCAAAG 850
```

```
                 .         .         .         .         .
 818 TGACGACGAAGAAGAAGTATTCGGAGGAATTGTTGATGAAGAGGACTGGA 867
     ||||||||||||||||||||||||||||||||||||||||||||||||||
 851 TGACGACGAAGAAGAAGTATTCGGAGGAATTGTTGATGAAGAGGACTGGA 900
                 .         .         .         .         .
 868 AACCTGATGACGATGATCCATCCGCTTGCGTAGTACCCGATAAAGTGAAC 917
     ||||||||||||||||||||||||||||||||||||||||||||||||||
 901 AACCTGATGACGATGATCCATCCGCTTGCGTAGTACCCGATAAAGTGAAC 950
                 .         .         .         .         .
 918 TTCTCTTCTGGAGCTGCCATTGATGTTGCAATGGTAAGTGTTTAGAATTT 967
     ||||||||||||||||||||||||||||||||||||||||||||||||||
 951 TTCTCTTCTGGAGCTGCCATTGATGTTGCAATGGTAAGTGTTTAGAATTT 1000
                 .         .         .         .         .
 968 ATCAATTACTCAACATTCGCTGATAAATAATTGGCTAATAAATAAAATTT 1017
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1001 ATCAATTACTCAACATTCGCTGATAAATAATTGGCTAATAAATAAAATTT 1050
                 .         .         .         .         .
1018 GAAATACATTACTTTTTACAGGTGCATAGTGCCGTTGAATTTATGACTGA 1067
      |||||||||||||||||||||||||||||||||||||||||||||||||
1051 AAAATACATTACTTTTTACAGGTGCATAGTGCCGTTGAATTTATGACTGA 1100
                 .         .         .         .         .
1068 TGTCAGAACAAAGAAACTACGATCTTTTGCTTCAATGCAGCGTAGGTATC 1117
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1101 TGTCAGAACAAAGAAACTACGATCTTTTGCTTCAATGCAGCGTAGGTATC 1150
                 .         .         .         .         .
1118 GTTTTATTAAAACGCAACATGACATGCAGAAACTTCGCGTTTTTGCTAAA 1167
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1151 GTTTTATTAAAACGCAACATGACATGCAGAAACTTCGCGTTTTTGCTAAA 1200
                 .         .         .         .         .
1168 AATAGTGAGTATTAACAGCTTCATATTCGGTATAAAACTGGGTTTTTAAG 1217
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1201 AATAGTGAGTATTAACAGCTTCATATTCGGTATAAAACTGGGTTTTTAAG 1250
                 .         .         .         .         .
1218 ACGAAATTCAATGCTCACGTGTTTCACAATTTTCGACACTTTCTGGACTT 1267
     ||||||||| ||||||||||||||||||||||||||||||||||||||||
1251 ACGAAATTCGATGCTCACGTGTTTCACAATTTTCGACACTTTCTGGACTT 1300
```

```
                 .         .         .         .         .
1268 CTTCGTACAAAAGTTTTTGAGGCAATCGATGACAGTGAGTATTCCTATTA 1317
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1301 CTTCGTACAAAAGTTTTTGAGGCAATCGATGACAGTGAGTATTCCTATTA 1350
                 .         .         .         .         .
1318 TTGAAAAACTACTGTGTTTGCACGACAAGTAGTGCATTCTTTGTCAGACT 1367
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1351 TTGAAAAACTACTGTGTTTGCACGACAAGTAGTGCATTCTTTGTCAGACT 1400
                 .         .         .         .         .
1368 TAAAACACATCTTGAAGGAATTTCGATTGACAAGTTCACGCTACGCCGTC 1417
     |||||||||| |||||||||||||||||||||||||||||||||||||||
1401 TAAAACACATTTTGAAGGAATTTCGATTGACAAGTTCACGCTACGCCGTC 1450
                 .         .         .         .         .
1418 TTGCAGTGCAATTGAATGATGAGCACGTCCATATTGAAGGATTTCAAGCA 1467
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1451 TTGCAGTGCAATTGAATGATGAGCACGTCCATATTGAAGGATTTCAAGCA 1500
                 .         .         .         .         .
1468 AGCGATGGCTGGCTGAAGAAGTGGAAAAAGACAAACGGTCTCGTTTCTCG 1517
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1501 AGCGATGGCTGGCTGAAGAAGTGGAAAAAGACAAACGGTCTCGTTTCTCG 1550
                 .         .         .         .         .
1518 CCACGTAACTACTTTCATCACTCGTGCCAATTACGTCAATAAAGAGCTCA 1567
     |||||||||||||||||||||||||||||| |||||||||||||||||||
1551 CCACGTAACTACTTTCATCACTCGTGCCAACTACGTCAATAAAGAGCTCA 1600
                 .         .         .         .         .
1568 CAGAACAAGCTGCCAAAAAGTTCGTGGAGGAAGTTAAAGCAGAATTGGCA 1617
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1601 CAGAACAAGCTGCCAAAAAGTTCGTGGAGGAAGTTAAAGCAGAATTGGCA 1650
                 .         .         .         .         .
1618 ACTTTGGATCCTGATGTCGTTTATAACTGTGACCAAAGTGGGTTCACGAA 1667
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1651 ACTTTGGATCCTGATGTCGTTTATAACTGTGACCAAAGTGGGTTCACGAA 1700
                 .         .         .         .         .
1668 AGAACAATATTGCAAACGGTAAATTCTAAACCGAGTTTTTCAAAGATTAT 1717
     |||||||||||||||||||||||||||||||||||||||||||||  ||
1701 AGAACAATATTGCAAACGGTAAATTCTAAACCGAGTTTTTCAAAGACCAT 1750
```

```
                .         .         .         .         .
1718 TAAAATTTTTAGGACGCTCGCACCAAAAGGTGTTAAACGTGTTGAAAGAC 1767
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1751 TAAAATTTTTAGGACGCTCGCACCAAAAGGTGTTAAACGTGTTGAAAGAC 1800
                .         .         .         .         .
1768 TGGTACAGTCCAAAGATGCCCTCACGCACTCTTACACAATCCTTCCCATG 1817
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1801 TGGTACAGTCCAAAGATGCCCTCACGCACTCTTACACAATCCTTCCCATG 1850
                .         .         .         .         .
1818 TTAAGCGCTTCCGGAAAGTTAGCCCCAATGTTGTACGTGGTTCTGCAGGT 1867
     ||||||||||||||||||||||||||||| ||||||||||||||||||||
1851 TTAAGCGCTTCCGGAAAGTTAGCCCCAAAGTTGTACGTGGTTCTGCAGGT 1900
                .         .         .         .         .
1868 ATGTTTGACAATATGCACAACATTGCCACACAGTCTTGTGACTATCGTTT 1917
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1901 ATGTTTGACAATATGCACAACATTGCCACACAGTCTTGTGACTATCGTTT 1950
                .         .         .         .         .
1918 TACATTATGCAACTTTATTAAATTGTAGGAGAAAGGTGGAAAATTTCCCA 1967
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1951 TACATTATGCAACTTTATTAAATTGTAGGAGAAAGGTGGAAAATTTCCCA 2000
                .         .         .         .         .
1968 AAAAAGGGCACTTCTCACCAGACAATCTGATCATCCGAGCTAATACGTCC 2017
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2001 AAAAAGGGCACTTCTCACCAGACAATCTGATCATCCGAGCTAATACGTCC 2050
                .         .         .         .         .
2018 CACATTATGAATAAACAACTAATGGTCGACTGGGTTGAATCCGCTGTTTG 2067
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2051 CACATTATGAATAAACAACTAATGGTCGACTGGGTTGAATCCGCTGTTTG 2100
                .         .         .         .         .
2068 TGATCCTTCGATGCCAACCGAGGTTGTCCTGCTTCTAGACGCTTGGCCTG 2117
     ||||||||||||||||||||||||||||| ||||||||||||||||||||
2101 TGATCCTTCGATGCCAACCGAGGTTGTCCAGCTTCTAGACGCTTGGCCTG 2150
                .         .         .         .         .
2118 CTTGGAAAAACGAAGGGGATGTTCAAGCTGCAGCATTATCCGGAAATACA 2167
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2151 CTTGGAAAAACGAAGGGGATGTTCAAGCTGCAGCATTATCCGGAAATACA 2200
```

```
                 .         .         .         .         .
2168 GTACATGTGAGATCTATTCCACCAGGAGCTACATCATTTATTCAACCTTG 2217
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2201 GTACATGTGAGATCTATTCCACCAGGAGCTACATCATTTATTCAACCTTG 2250
                 .         .         .         .         .
2218 CGATCTTTACTTTTTCTGTCCGTTGAAGAATTTTGTCAAAAAGGTGAACG 2267
     ||||||||||||||||||||||||||||||||| ||||||||||||||||
2251 CGATCTTTACTTTTTCTGTCCGTTGAAGAATTTCGTCAAAAAGGTGAACG 2300
                 .         .         .         .         .
2268 CGTACATCATCTACTCCGGTATCACCTTCAAGACGTCAGAGCGTGACAAC 2317
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2301 CGTACATCATCTACTCCGGTATCACCTTCAAGACGTCAGAGCGTGACAAC 2350
                 .         .         .         .         .
2318 CTGCTTCGCGTGATATCTGCAGTGTACCGTGTCTTTCGTGCACCAATTTT 2367
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2351 CTGCTTCGCGTGATATCTGCAGTGTACCGTGTCTTTCGTGCACCAATTTT 2400
                 .         .         .         .         .
2368 CCAATCATGCTGGAAGTACGGCTGGATCCAAGGAGGATACATAGATGACC 2417
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2401 CCAATCATGCTGGAAGTACGGCTGGATCCAAGGAGGATACATAGATGACC 2450
                 .         .         .         .         .
2418 AACATGTCAAAGTGGAAACTCCATCCAAATTTTGTTTCAAAGTTTCTGGA 2467
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2451 AACATGTCAAAGTGGAAACTCCATCCAAATTTTGTTTCAAAGTTTCTGGA 2500
                 .         .         .         .         .
2468 TACTGTTCGCAAAAGAAAACGAGAGATACGATGTGTCAAGATACGGCTTT 2517
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2501 TACTGTTCGCAAAAGAAAACGAGAGATACGATGTGTCAAGATACGGCTTT 2550
                 .         .         .         .         .
2518 TCTTCTTTGCCCATACTGTAAGAAGGTTTTATGCTTTAACCACTGGGTTG 2567
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2551 TCTTCTTTGCCCATACTGTAAGAAGGTTTTATGCTTTAACCACTGGGTTG 2600
                 .         .         .         .         .
2568 GATGCGGCTTCCCAGCTCATAAGTGTAAGTGTTAAAAGCCATTGTTGAGT 2617
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2601 GATGCGGCTTCCCAGCTCATAAGTGTAAGTGTTAAAAGCCATTGTTGAGT 2650
```

```
                 .         .         .         .         .
2618 ATATTATATGTTGCTTTTGTTTTTTTTTTTAATATTGGCATCGTTCGTTT 2667
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2651 ATATTATATGTTGCTTTTGTTTTTTTTTTTAATATTGGCATCGTTCGTTT 2700

                 .         .         .         .         .
2668 GTTTTTTACATAAACTTTAAACATCTGTTTTATTGTGCTTTTAACTATAA 2717
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2701 GTTTTTTACATAAACTTTAAACATCTGTTTTATTGTGCTTTTAACTATAA 2750

                 .         .         .         .         .
2718 CACATATTGA.AAAAAAAAATCGGAAAAATTTGGAGAATTGACGCATTTG 2766
     |||||||||| |||||||||||||||||||||||||||||||||||||||
2751 CACATATTGAGAAAAAAAAAATCGGAAAAATTTGGAGAATTGACGCATTTG 2800

                 .         .         .         .         .
2767 ATTCATAGGGATAATATCTGCACATTTTCCCACTTGAAAAAGTTTATAGA 2816
     |||||||||||||||||||||| |||||||||||||||||||||||||||
2801 ATTCATAGGGATAATATCTGCATATTTTCCCACTTGAAAAAGTTTATAGA 2850

                 .         .         .         .         .
2817 TGTGTTCTAATTGGAAATGGATTGACTTTAAAAATTAAAAAAAAAAGTTT 2866
     |||||||||||||||||||||||||||||||||||| |||||||||||||
2851 TGTGTTCTAATTGGAAATGGATTGACTTTAAAAATT.AAAAAAAAAGTTT 2899

                 .         .         .         .         .
2867 ATGTTGACGGGGAACCGAACCACGAATACCACACCTCGAAACTCACGCCA 2916
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2900 ATGTTGACGGGGAACCGAACCACGAATACCACACCTCGAAACTCACGCCA 2949

                 .         .         .         .         .
2917 CTACCACTCGGCCACCAAGCCATGTTTGCTAACGCTAATTGAGAGTGGTG 2966
     ||||||||||||||||||||||||||||||||||||||||||||||||||
2950 CTACCACTCGGCCACCAAGCCATGTTTGCTAACGCTAATTGAGAGTGGTG 2999

                 .         .         .         .         .
2967 AGTTCACTGGTGTTTCTGCAGCGTGGTAGTCATTGAAATATCACTTGATA 3016
     ||||||||||||||||||||||||||||||||||||||||||||||||||
3000 AGTTCACTGGTGTTTCTGCAGCGTGGTAGTCATTGAAATATCACTTGATA 3049

                 .         .         .         .         .
3017 TGAGGAATTTCAGCACCAAAATTCAAGACGTCACAACTTGGCCAAACCTG 3066
     |||||||||||||||||||| |||||||||||||||||||||||||||||
3050 TGAGGAATTTCAGCACCAAACTTCAAGACGTCACAACTTGGCCAAACCTG 3099
```

```
                .         .         .         .         .
3067 CTCTGAAGCAGCTAAATTTTTTGAAGGGATCATCTGTAACTATCCTCTAT 3116
     |||||||||||||||||||||| |||||||||||||||||||||||||||
3100 CTCTGAAGCAGCTAAATTTTTTGGAGGGATCATCTGTAACTATCCTCTAT 3149
                .         .         .         .         .
3117 CGAATGTATATAGTGGCGTAACGAAGTCCGATAACGAGTTCAGAACCTTC 3166
     ||||||||||||||||||||||||||||||||||||||||||||||||||
3150 CGAATGTATATAGTGGCGTAACGAAGTCCGATAACGAGTTCAGAACCTTC 3199
                .
3167 CCTTG..... 3171
     |||||
3200 CCTTGTTAGGTGAAC 3214
```

APPENDIX G: Alignment of four cosmid sequences identified as high scoring blast hits to Tc5.

```
         1                                                                                                                  100
  C01b7  .......TTC ACACTTACAA GGGAAGTCTT TGAGGGGGTC CGTAGATTTG GGGTTCTCAT GCTAAAATTC CTACAGAAGA GTGTTAGTTA TGATCTCTCC
  T14g8  .......TAA ACGTTTTCAA GGGAAGTCTT TGAGGGGGTC CGTAGATTTG GGGTTCTCAT GCTAAAATTC CTACAGAAGA GTGTTAGTTA TGATCTCTCC
T19d7rc  .......AGA TACCTAACAA GGGAAGTCTT TGAGGGGGTC CGTAGATTTG GGGTTCTCAT GCTAAAATTC CTACAGAAGA GTGTTAGTTA TGATCTCTCC
C48b4rc  .......ATG CAACTGACAA GGGAAGTCTT TGAGGGGGTC CGTAGATTTG GGGTTCTCAT GCTAAAATTC CTACAGAAGA GTGTTAGTTA TGATCTCTCC

         101                                                                                                                200
  C01b7  AAAAAATTTA GCTGCCCCGG TCAAGTTTCA GCAAAGTTAT GACGTTTTTA AATTTCAGTT AAAAACACCA TTGAAATCCA CTGTCTTACC ATGCAATCCA
  T14g8  AAAAAATTTA GCTGCCCCGG TCAAGTTTCA GCAAAGTTAT GACGTTTTGA AATTTCAGTT AAAAACACCA TTGAAATCCA CTGTCTTACC ATGCAATCCA
T19d7rc  AAAAAATTTA GCTGCCCCGG TCAAGTTTCA GCAAAGTTAT GACGTTTTGA AATTTCGGTT AAAAACACCA TTGAAATCCA CTGTCTTACC ATGCAGTCCA
C48b4rc  AAAAAAATTA GCTGCCCAGG TCAAGTTTCA GCAAAGTTAT GACGTTTTGA AATTTCAGTT AAAAACACCA TTGAAATCCA CTGTCTTACC ATGCAATCCA

         201                                                                                                                300
  C01b7  CGCAAATCTC AGCTTGCGTG ACCACCGAAA ATGTGACACC CAC.CACATT GAGTTGAAAA ATGTCCTCGG TGGCCGAG.. .......... TTGGGAGTGC
  T14g8  CGCAAATCTC AGCTTGCGTG ACCACCGAAA ATGTGACACC CAC.CACATT GAGTTGAAAA ATGTCCTCGG TGGCCGAG.. .......... TTGGGAGTGC
T19d7rc  CGCAAATCTC AGCTTGCGTG ACCACCGAAA ATGTGACACC CAT.CACATT GAGTTGAAAA ATGTCCTCGG TGGACGAGTT AACTCCCCAA TTGGGAGTGC
C48b4rc  CGCAAATCTC AGCTTGCGTG ACCACCGAAA ATGTGACACC CACTCACATT GAGTTGAAAA ATGTCCTCGG TGGCCGAG.. .......... TTGGGAGTGC

         301                                                                                                                400
  C01b7  GCGGGTCTGA TAAGATTTAA GCTTTGGTTC GATTCCTTCT ATTTTTGAAA TATTTTTGTA AGTTGAATAA AGTTGTAAAA CAACTCATTC AAACATTTTT
  T14g8  GCGCGTCTGA TAAGATTTAA GCTTTGGTTC GATTCCTTCT ATTTTTGAAA TATTTTTGTA AGTTGAATAA AGTTGTAAAA CAACTCATTC AAACATTTTT
T19d7rc  GCGCGTCTGA TAAGATTTAA GCTTTGGTTC GATTCCTTCT ATTTTTGAAA TATTTTTGTA AGTTGAATAA AGTTGTAAAA CAACTCATTC AAACATTTTT
C48b4rc  GCGCGTCTGA TAAGATTTAA GCTTTGGTTC GATTCCTTCT ATTTTTGAAA TATTTTTGTA AGTTGAATAA AGTTGTAAAA CAACTCATTC AAACATTTTT

         401                                                                                                                500
  C01b7  GCGCATTTTT AAAGTGATTT TATTCTTATT CGGGAACCTA GAATCATTGT CCGCACTTTT TAGAAATTTT TATTTTTTTC ATTTTTACTC AAAATTTCTT
  T14g8  GCGCATTTTT AAAGTGATTT TATTCTTATT CGGGAACCTA GAATCATTGT CCGCACTTTT TGGAAATTTT TATTTTTTTC ATTTTTGTTC AAAATTTCTT
T19d7rc  GCGCATTTTT AAAGTGATTT TATTCTTATT CGGGAGCCTA GAATCATTGT CCGCACTTTT TGGAAATTTT TATTTTTTTC ATTGTTGCTC AAAATTTCTT
C48b4rc  GCGCATTTTT AAAGTGATTT TATTCTTATT CGGGAGCCT. .......... .......... .......... .......... .......... ..........
```

```
         501                                                                                                                  600
  C01b7  GATCAACTCC AAGCAAAAAA ATAAAAAAAT TTCATTTTTC TAAACAATTA TGAAATTGCT ATGTTGTTGT TCAGAAATGT ATGAAACGTA CATTACACAA
  T14g8  GATCAACTCC AAGCAAAAAA ATCAAAAAAT TTCATTTTTC TAAACAATTA TGAAATTGCT ATGTTGTTGT TCAGAAATGT ATGAAACGTA CATTACACAA
T19d7rc  GATCAACTCC AAGCAAAAAA TTCAAAAAAT TTCATTTTTC TAAACAATTA TGAAATTGCT ATGTTGTTGT TCAGAAATGT ATGAAACGTA CATTACACAA
C48b4rc  .......... .......... .......... ...ATTTTTC TAAACAATTA TGAAATTGCT ATGTTGTTGT TCAGAAATGT ATGAAACGTA CATTACACAA

         601                                                                                                                  700
  C01b7  GTTTTAACTC TCTATTCGCA AGTAAACCGT CGAAATGATC TACATCTCAC GAACTTTGTG CAAAATATTT AACCAACTTT GAAGTTGCAT AACTTCGTTG
  T14g8  GTTTTAACTC TCTATTCGCA AGTAAACCGT CGAAATGATC TACATCTCAC GAACTTTGTG CAAAATATTT AACCAACTTT GAAGTTGCAT AACTTCGTTG
T19d7rc  GTTTTAACTC TCTATTCGCA AGTAAACCGT CGAAATGATC TACATCTCAC GAACTTTGTG CAAAATATGT AACCAACTTT GAAGTTGCAT AACTTCGTTG
C48b4rc  GTTTTAACTC TCTATTCGCA AGTAAACCGT CGAAATGATC TACCTCTCAC GAACTTTGTG CAAAATATTT AACCAACTTT GAAGTTGCAC AACTTCGTTG

         701                                                                                                                  800
  C01b7  AGATAAATTA TTTTGAAAAA TGATCACCCA ACAAAATGTT TGTTGAATAA CAGTGAACAA AGTTTTAGTT ATAAACTTTT TGATACCTCC AGCTACAAAG
  T14g8  AGATAAATTA TTTTGAAAAA TGATCAACTA ACAAAATGTT TGTTGAATAA CAGTGAACAA AGTTTTAGTT ATAAACTTTT TGATACCTCC AGCTACAAAG
T19d7rc  AGATAAATTA TTTTGAAAAA TGATCAACTA ACGAAATGTT TGTTGAATAA CAGTGAACAA AGTTTTAGTT ATAAACTTTT TGATACCTCC AGCTACAAAG
C48b4rc  AGATAAATTA TTTTGAAAAA TGATCAACTA ACAAAATGTT TGTTGAATAA TAGTGAACAA AGTTTTAGTT ATAAACTTTT TGATACCTCC AGCTACAAAG

         801                                                                                                                  900
  C01b7  AAGAAAACAA GGTTGGCATT TGGCTAGTTT TTCTATTAAC ATTGTGTTTT GGAAAACGGT CACAACTTTT TGGTGGCTGA AGGTATCAAA AAGTTTATAA
  T14g8  AAGAAAACAA GGTTGGCATT TGGCTAGTTT TTCTATTAAC ATTGTGTTTT GGAAAACGGT CACAACTTTT TGGTGGCTGA AGGTATCAAA AAGTTTATAA
T19d7rc  AAGAAAACAA GGTTGGCATT TGGCTAGTTT TTCTATTAAC ATTGTGTTTT GGAAAACGGT CACAACTTTT TGGTGGCTGA AGGTATCAAA AAGTTTATAA
C48b4rc  AAGAAAACAA GGTTGGCATT TGGCTAGTTT TTCTATTAAC ATTGTGTTTT GGAAAACGGT CACAACTTTT TGGTGGCTGA AGGTATCAAA AAGTTTATAA

         901                                                                                                                 1000
  C01b7  CTAAAACTTT GTTCACTGTT ATTCAACAAA CATTTTGTTA GTTGATCATT TTTCAAAATA ATTTATCTCA ACGAAGTTA. TGCAACTTCA AAGTTGGTTA
  T14g8  CTAAAACTTT GTTCACTGTT ATTCAACAAA CATTTTGTTA GTTGATCATT TTTCAAAATA ATTTATCTCA ACGAAGTTA. TGCAACTTCA AAGTTGGTTA
T19d7rc  CTAAAACTTT GTTCACTGTT ATTCAACAAA CATTTTGTTA GTTGATCATT TTTCAAAATA ATTTATCTCA ACGAAGTTA. TGCAACTTCA AAGTTGGTTA
C48b4rc  CTAAAACTTT GTTCACTATT ATTCAACAAA CATTTTGTTA GTTGATCATT TTTCAAAATA ATTTATCTCA ACGAAGTTAC TGCAACTTCA AAGTTGGTTA

         1001                                                                                                                1100
  C01b7  AATATTTTGC ACAAAGTTCG TGAGATGTAG ATCATTTCGA CGGTTTACTT GCGAATAGAG AGTTAAAACT TGTGTAATGT ACGTTTCATA CATTTCTGAA
  T14g8  AATATTTTTC ACAAAGTTCG TGAGATGT.. ATCATTTCGA CGGTTTACTT GCGAATAGAG AGTTAAAACT TGTGTAATGT ACGTTTCATA CATTTCTGAA
T19d7rc  CATATTTTGC ACAAAGTTCG TGAGATGTAG ATCATTTCGA CGGTTTACTT GCGAATAGAG AGTTAAAACT TGTGTAATGT ACGTTTCATA CATTTCTGAA
C48b4rc  AATATTTTGC ACAAAGTTCG TGAGATGTAG ATCATTTCGA CGGTTTACTT GCGAATAGAG AGTTAAAACT TGTGTAATGT ACGTTTCATA CATTTCTGAA
```

```
         1101                                                                                                                  1200
  C01b7  CAACAACATA GCAATTTCAT AATTGTTTAG AAAAATGAAA TTTTTTGATT TTTTTGCTTG GAGTTGATCA AGAAATTTTG AGCAAAAATG AAAAAAATAA
  T14g8  CAACAACATA GCAATTTCAT AATTGTTTAG AAAAATGAAA TTTTTTGATT TTTTTGCTTG GAGTTGATCA AGAAATTTTG AACAAAAATG AAAAAAATAA
T19d7rc  CAACAACATA GCAATTTCAT AATTGTTTAG AAAAATGAAA TTTTTTGAAT TTTTTGCTTG GTGTTGATCA AGAAATTTTG AGCAACAATG AAAAAAATAA
C48b4rc  CAACAACATA GCAATTTCAT AATTGTTTAG AAAAA..... .......... .......... .......... .......... .......... ..........

         1201                                                                                                                  1300
  C01b7  AAATTTCCAA AAAGTGCGGA CAATGATTCT AGGTTCCCGA ATAAAAATAA AATCACTTTA AAAATGCGCA AAAATGTTTG AATGAGTTGT TTTACAATTT
  T14g8  AAATTTCCAA AAAGTGCGGA CAATGATTCT AGGTTCCCGA ATAAGAATAA AATCACTTTA AAAATGCGCA AAAATGTTTG AATGAGTTGT TTTACAACTT
T19d7rc  AAATTTCCAA AAAGTGCGGA CAATGATTCT AGGTTCCCGA ATAAGAATAA AATCACTTTA AAAATGCGCA AAAATGTTTG AATGAGTTGT TTTACAACTT
C48b4rc  .......... .......... .........T AGGCTCCCGA ATAAGAATAA AATCACTTTA AAAATGCGCA AAAATGTTTG AATGAGTTGT TTTACAACTT

         1301                                                                                                                  1400
  C01b7  TATTCAACTT ACAAAAATAT TTCAAAAATA GAAGGAATCG AACCAAAGCT TAAATCTTAT CAGACGCGCG CACTCCC... .........A ACTCGGCCAC
  T14g8  TATTCAACTT ACAAAAATAT TTCAAAAATA GAAGGAATCG AACCAAAGCT TAAACCTTAT CAGACGCGCG CACTCCC... .........A ACTCGGCCAC
T19d7rc  TATTCAACTT ACAAAAATAT TTCAAAAATA GAAGGAATCG AACCAAAGCT TAAATCTTAT CAGACGCGCG CACTCCCAAT TGGGGAGTTA ACTCGTCCAC
C48b4rc  TATTCAACTT ACAAAAATAT TTCAAAAATA GAAGGAATCG AACCAAAGCT TAAATCTTAT CAGACGCGCG CACTCCC... .........A ACTCGGCCAC

         1401                                                                                                                  1500
  C01b7  CGAGGACATT TTTCAACTCA ATGTG.GTGG GTGTCACATT TTCGGTGGTC ACGCAAGCTG AGATTTGCGT GGATTGCATG GTAAGACAGT GGATTTCAAT
  T14g8  CGAGGACATT TTTCAACTCA ATGTG.GTGG GTGTCACATT TTCGGTGGTC ACGCAAGCTG AGATTTGCGT GGATTGCATG GTAAGACAGT GGATTTCAAT
T19d7rc  CGAGGACATT TTTCAACTCA ATGTG.ATGG GTGTCACATT TTCGGTGGTC ACGCAAGCTG AGATTTGCGT GGACTGCATG GTAAGACAGT GGATTTCAAT
C48b4rc  CGAGGACATT TTTCAACTCA ATGTGAGTGG GTGTCACATT TTCGGTGGTC ACGCAAGCTG AGATTTGCGT GGATTGCATG GTAAGACAGT GGATTTCAAT

         1501                                                                                                                  1600
  C01b7  GGTGTTTTTA ACTGAAATTT CAAAACGTCA TAACTTTGCT GAAACTTGAC CGGGACAGCT AAATTTTTTG GAGAGATCAT AACTAACACT CTTCTATCGG
  T14g8  GGTGTTTTTA ACTGAAATTT CAAAACGTCA TAACTTTGCT GAAACTTGAC CGGGGCAGCT AAATTTTTTG GAGAGATCAT AACTAACACT CTTCTGTAGG
T19d7rc  GGTGTTTTTA ACTGAAATTT CAAAACGTCA TAACTTTGCT GAAACTTGAC CGGGGCAGCT AAATTTTTTG GAGAGATCAT AACTAACACT CTTCTGTAGG
C48b4rc  GGTGTTTTTA ACTGAAATTT CAAAACGTCA TAACTTTGCT GAAACTTGAC CTGGGCAGCT AAATTTTTTG GAGAGATCAT AACTAACACT CTTCTGTAGG

         1601                                                    1663
  C01b7  AATTTCAATA TGAGAACCCC AAATCTACGG A.CCCCTCAA AGACTTCTCT TGTTAAGTTT TTG
  T14g8  AATTTTAGCA TGAGAATCCC AAATCTACGG ACCCCCTCAA AGACTTCCCT TGTTAGGTCG TA.
T19d7rc  AATTTTAACA TGAGAACCCC AAATCTACGG ACCCCCTCAA AGACTTCCCT TGTAAAGCAA AC.
C48b4rc  AATTTTAGCA TGAGAACCCC AAATCTACGG ACCCCCTCAA AGACTTCCCT TGTGAGTGTT AA.
 C39d10  AAAAATACTA TGTCATGCCA AAGTCAGACA TTAAGTTTAT CGACTTCCCT TGTAAAGCTC CA.
```

APPENDIX H: Alignment of five short cosmid sequences identified as high scoring blast hits to Tc5.

```
         1                                                                                                                  100
  C24a3  .TCATGACTG ACAAGGGAAG GCTCTGAATT CGTTATCGGA GTTCGTTACG CCACTGTATA CATTCGATAG AGAATGGTTA CAGATGATCA CTCCAAAAAA
F44b9rc  .TTGATGCTT ACAAGGGAAG GCTCTGAACT CGTTATCGGA CTTCGTTACG CCACTGTATA AATTCGATAG AGAATGGTTA CAGATGATCA CTCCTAAAAA
  C04e7  .AAAGCCCTA ACAAGGGAAG GCTCTGAATT CGTTATCGGA CTTCGTTACG CCACTGTATA CATTCGATAG AGAATGGTTA CAGATGATCA CTCCAAAAGA
Zk930rc  TCTAAACCTA ACAAGGGAAG GCTCT..... .......GGA CTTCGTTACG CCACTGTATA CATTCGATAG AGAATGGTTA CAGATGATCA CTCCAAAAAA
 C39d10  ATATTTTTTA AAGATGTAAG TTGAAAAATT TAATGTCTGA CTTTGGCATG AAATAGTATT TTTTCGATAA AATAATGAAA ATAATGATCC CTCCAAAAAA

         101                                                                                                                200
  C24a3  TTTAGCTGCT TCAGAGCAGG TTCGACCAAG TTACGACACT TTGAAGTTAC CGAAAAAAAA ATCCTTGATG CCCCCTTTGC CCCCTTTGAA CCCCCTTTGA
F44b9rc  TTTAGCTGCT TCAGAGCAGG TTCGACCAAG TTACGACACT TTGAAGTTGC CGAAAAAAAA .......... .......... .......... ...TCCTTGA
  C04e7  TTTACCTGCT TTAGAGCAGG TTCGACCAAG TTACGACACT TTAAAGTTGC CGAAAAAAAA TCCTTGAAGA CC........ .......... ..........
Zk930rc  TTTAGCTGCT TCAGAGCAGG TTCGACCAAG TTACGACACT TTAAAGTTGC CGACAAAAAA .......... .......... .......... ...TCCTTCA
 C39d10  TTTAGCTGCC CCGATCCAGG TTCAGCAAAG TTATGACGTT TTGAAAGTGA CTAAAACA.. .......... .......... .......... ..........

         201                                                                                                                300
  C24a3  AAAAACCCCT TTGAAAAAAA TCTAAAATTT TCACTGAAAA ATTGTTTTTC TGAAAGTTGA TAAAAATAGT TGTAATCGAT TTAAAATAGT .AAAA....A
F44b9rc  TGCACCCCCT TTGAAAAAAA TCTAAAATTT TCACTGAAAA ATTGTTTTTC TGAAAGTTGA TAAAAATAGT TGTAATCGAT TTAAAATAGT .TAAA....A
  C04e7  .TCTGAACCT TTGAAAAAAA TCTAAAATTT CCACTGAAAA ATTGTTTTTC TGCAAGTTGA T.AAAATAGT TGTAATCGAT TTAAAATAGT AAAAA....A
Zk930rc  TGCACCCCCT TTGAAAAAAA TCTAAAATTT TCACTAAGAA TTTTTTTTTC TGAAAGTTGA TAAAAATAGT TGTAATCGAT TTAAAACAGT AAAAAACATA
 C39d10  ......CCTT TTTACAAAAT TTCAAAATTT TCAAAAAAAA AACATTTTTT TCTAAAAAGA GGAAAAAATG TTTGCAAGTT .......... ..........

         301                                                                                                                400
  C24a3  ACATATATTA TACACGTTTT AGCTCATCAA TCTC.AAAAA AACCCTTAAA ATAATCTACA TATCCTGAGA AAAATTCCAA AAAGTAGATG TTCATGTAGA
F44b9rc  ACATATATTA TACAAGTTTT AGCCCATCAC TCTC.AAAAA AACCCTTAAA TTAATCTACA TATCTTGAGA AAAATTCCAA AAAGTAGATG TTCATGTAGA
  C04e7  ACATATATTA TACAAGTTTT AGCCCATCAC TCTCAAAAAA AACCCTTAAA ATAATCTACA TATCCTGAGA AAAATACCAA AAAGTAGATG TTCATGTAGA
Zk930rc  TAAAACATTA TACAAGTTTT AGCCCGTCAC TCTC.AAAAA AACCCTTAAA ATAATCTACA CATCTTGAGA AAAATTCCAA AAAATAGATG TTCATGTAGA
 C39d10  .......... .....GTTTT AGCTCACAAC TCTC.AAGAA AACCCACAAA CTAATCTA.. .......... .......CCA GAAAAAGTAC TTTTTGGAAT

         401                                                                                                                500
  C24a3  TCAATTCAAG CG......TT TTTTGAGAAT AATGAACTGA AACTTGTATG GTATGATTTT TCTATCATTT CCAACTGTCT GAAAAC.GTT TATATAAACT
F44b9rc  TCAATTTAAG GG.....TTT TTTTGAGAAT TATGAACTGA AACTTGTATG GTATGATTTT TCCATCATTT TCAACTATTT GAAAACATTT TATATCAACT
  C04e7  TCAATTCAAG GA.....TTT TTTTGAGAAT TATGAACTGA AACTTGTATG GTATGATTTT TCCATTATTT CCAACTATTT GAAAAC.GTT TAAATCAACT
Zk930rc  TCAATTCAAG GG..TTTTTT TTTTGAGAAT TATGATCTGA AACTTGTATG GTATGATTTT TCCATCATTT CCAACCATTT GAAAAC.GTT TATATCAAGT
 C39d10  TTTGCTCAAG GTATGTGGTT TTTTGTCCGT GGTGAGAAAA AACGTGTATG ATATATGCTT TTCACTGTTT TGGA.TAAGT TAAAACAGTT TTTATCGATT
```

```
         501                                                                                                                  600
  C24a3  TTTAGAAAAA CAATTTTTTT TTGAATTTTT TGGAATTTTT .......... ..AAAAAGGG GTATTTTGGC ACTTTCAAAA TGTCGTAACC TGCTCTGAAG
F44b9rc  TTTAGAAAAA AAATTTTTTT T..AATTTTT TGGAATTTTT .......... ..AAAAAGGG GTATTTTGGC ACTTTCAAAA TGTCGTAACC TGCTCTGAAA
  C04e7  TTTCGAAAAA AATTTTTTTT .......... ..GAATTTTT .......... ..AAAAAGGG GTATTTTGGC ACTTTCAAAA TGTCGTAACT TGGTCGAACC
Zk930rc  TTTAGAATAA AAAAAAATT. ..GAATTTTT TGGAATTTTT .......... ..AAAAAGGA GTATTTTGGC ACTTTCAAAA TGTCGTAACT TGGTCGAACC
 C39d10  TTCAGAAAAC AATTTTTTTT TTTGAAATTT TAAAAGTTTT TTTTTTGGTA GGTCAAGAAA TTTTTTCGGC ACTTTCAAAA CATCATAACT TTGCTAAACC

         601                                                                                                                  700
  C24a3  TCGTCGTAAC CTGCTCTGAA GCAGCAAAAT TTTTGGAGTG ATCATCTGTA ACTATTCTCT ATCGAATGTA TATAGTGGCG TAACGAAGTC CGATAACGAG
F44b9rc  GCAGC..... .......... ....TAAATG TTTTGGATTG ATCATCTGTA ACCATTCTCT ATCGAATGTA TATAATGGCG TAACGAAGTC CGATAACGAG
  C04e7  TGCTCTGAAG CAGC...... ....TAAATT TTTTGGAGTG ATCATCTGTA ACTATTCTCT ATCGAATGTA TATAGTGGCG TAACAAAGTC CAATAACGAG
Zk930rc  TGCTCTGAAG CAGC...... ....TAAATT TTTTGGAGTG ATCATCTGTA ACTATTCTCT ATCGAATGTA TATAGTGGCG TAACGAAGTC CGATAAAAAT
 C39d10  TGGACCGGGG CAGC...... ....TAAATT TTTTGGAGGG ATCATTATTA TCACTCTTTT ATCGAAAAAA TACTATGTCA TGCCAAAGTC AGACATTAAG

         701                       728
  C24a3  TTCAGAGCCG TCCCTTGTGA GATTAAA.
F44b9rc  TTCAGAGCCT TCCCTTATTA GGTAAC..
  C04e7  TTCAGAGCCT TCATTTGAGC CTTCTTT.
Zk930rc  TTCCTGTGAT CTTCGCAACT TTCTCAGT
 C39d10  TTTATCGACT TCCCTTGTAA AGCTCCA.
```

APPENDIX I: Alignment of Tc6 and nine cosmid sequences identified as high scoring blast hits to Tc6.

```
           1                                                                                                         100
    T26a8  ..ATTTAAAC TACAGTGCTC CACATAATGA TACGGCCACC CCCAAATTTT AGTAAAACTC AAAACTGGGT TGAGATAGCA AAACATAGTT TCTTGTGAAA
    Zc395  ..TTTTAGAC TACAGTGCTC CACATAATGA TACGGCCACC CCCAAATTTT GGTATAACTC AAAACTGGGT TGAGATAGCA AAACATAGTT TCTTGTGAAA
    C33h5  ATGTTTAA.. TACAGTGCTC CACATAATGA TACGGCCACC CCCAAATTTT GGTACAACTC AAAACTGGGT TGAGATAGCA AAACATAGTT TCTTGTGAAA
    F48e8  ..GCTAAATG TACAGTGCTC CACATAATGA TACGGCCACC CCCAAATTTT GGTATAACTC AAAACTGGGT TGAGATAGCA AAACATAGTT TCTTGTGAAA
    Tc6.1  .......... ..CAGTGCTC CACATAATGA TACGGCCACC CCCAAATTTT GGTATAACTC AAAACTGGGT TGAGATAGCA AAACATAGTT TCTTGTGAAA
  Zk669rc  ..TGTAAATC TACAGTGCTC CACATAATGA TACGGCCACC CCCAAATTTT GGTATAACTC AAAACTGGGT TGAGATAGCA CAACATAGTT TCTTGTGAAA
  Zk180rc  ..TTTTGAAG TACAGTGCTC CACATAATGA TACGGCCACC CCCAAATTTT GGTATAACTC AAAACTGGGT TGAGATAGCA AAACATAGTT TCTTGTGAAA
    W03a3  ..TCAACTTG TACAGTGCTC CACATAATGA TACGGCCACC CCCAAATTTT GGTATAACTC AAAACTGGGT TGAGATAGCA AAACATAGTT TCTTGTGAAA
    F53b7  ..TCCTTTAC TACAGTGCTC CACATAATGA TACGGCCACC CCCAAATTTT GGTATAACTC AAAACTGGGT TGAGATAGCA AAACATAGTT TCTTGTGAAA
      Ac3  ..TATCTATA TACAGTGCTC CACAAAATGA TACGGCCACC CCTAAATTTT GGTATAACTC AAAACTGGGT TAAGATAGCA AAACATAGTT TCATGTGAAA

           101                                                                                                       200
    T26a8  ATGTTCGCTG TACTAACTTA CTTTCAGATA AGTATTGGAA ATATACCTGA ACCGTTCAGA AAAA..AGAT AAACCA.TTT TTTCATGAAA AACCATATTA
    Zc395  ATGTTCGCTG TACTGGCTAA CTTTCAGATA AGTATTGGAA ATATACCTGA ATCGTTCAGA AAAA..AGAC AAACCATTTT TTTCATGAAA AACCATATTA
    C33h5  ATGTTCGCTG TACTGGCTAA CTTTCAGATA AGTATTGGAA ATATACCTGA ACCGTTCGT. AAAA..AAGA TAAACCATTT TTTCATGAAA AACCATATTA
    F48e8  ATGTTCGCTG TACTGGCTAA CTTTCAGTTA AGTATTGGAA ATATACCTGA ACCGTTTGT. AAAA..AATA TAAACCATTT TTTCATGAAA AACCATA.TA
    Tc6.1  ATGTTCGCTG TACTGGCTAA CTTTCAGATA AGTATTGGAA ATATACCTGA ACCGTTCGTA AAAA..AAGA TAAACCATTT TTTCATGAAA AACCATATAA
  Zk669rc  ATGTTCGCTG TACTGGCTAA CTTTCAGATA AGTATTGGAA ATATACCTGA ACCGTTCGTA AAAA..AAGA TAAACCATTT TTTCATGAAA AACCATATAA
  Zk180rc  ATGTTCGCTG TACTGGCTAA CTTTCAGATA AGTATTGGAA ATATACCTGA ACCGTTCGTA AAAA..AAGA TAAACCATTT TTTCATGAAA AACCATATAA
    W03a3  ATGTTCGCTG TACTGGCTAA CTTTCAGATA AGTATTGGAA ATATACCTGA ACCGTTCGTA AAAA..AAGA TAAACCATTT TTTCATGAAA AACCATAT.A
    F53b7  ATGTCCGCTG TACTGGCTAA CTTTCAGATA AGTATTGGAA ATATACATGA ACCGTTCGT. AAAA..AAGA TAAACCATTT TTTCATGAAA ATCCATATAA
      Ac3  ATGTTCGTTG TACTGGCTAG CTTTGAGATA AGTGATGGAA ATATACCCGA AGCGTTCGTA AAAAAGAAAA AAAAACATTT TTTCATGAAA AACTATTTAA

           201                                                                                                       300
    T26a8  AAAAATCCAC AAAATGATAC GGCCACCCTT GGTTTTTGTT TTCTTTTTTC GTTTTCTTTG CAATTTTTTT TGCTAAACGT TAAGTTTCAT GTTCGTTTGT
    Zc395  AAAAATCCAC AAAATGATAC GGCCACCCTT GGTTTTCGTT TTCTTTTTTC GTTTTCTTTG CAATTTTTTT TGCTAAACGT TAAGTTTCAT GTTCGTTTGT
    C33h5  AAAAATCCAC AAAATGATAC GGCCACCCTT GGTTTTTGTT TTCTTTTTTC GTTTTCTTTG CAA.TTTTTT TGCTAAACGT TAAGTTTCAT GTTCGTTTGT
    F48e8  AAAAATCCAC AAAATGATAC GGCCACCCTT GGTTTTCGTT TTCTTTTTTC GTTTTCTTTG CAATTTTTTT TGCTAAACGT TAAGTTTCAT GTTCGTTTGT
    Tc6.1  AAAAATCCAC AAAATGATAC GGCCACCCTT GGTTTTTGTT TTCTTTTTTC GTTTTTTTTG CAATTTTTTT TGCTAAACGT TAGGTTTCAT GTTCGTTTGT
  Zk669rc  AAAAATCCAC AAAATGATAC GGCCACCCTT GGTTTTTGTT TTCTTTTTTC GTTTTTTTTG CAATTTTTTT TGCTAAACGT TAGGTTTCAT GTTCGTTTGT
  Zk180rc  AAAAATCCAC AAAATGATAC GGCCACCCTT GGTTTTTGTT TTCTTTTTTC GTTTTTTTTG CAATTTTTTT TGCTAAACGT TAGGTTTCAT GTTCGTTTGT
    W03a3  AAAAATCCAC AAAATGATAC GGCCACCCTT GGTTTTTGTT TTCTTTTTTC GTTTTTTTTG CAATTTTTTT TGCTAAACGT TAGGTTTCAT GTTCGTTTGT
    F53b7  AAAAATCCAC AAAATGATAC GGCCACCCTT GGTTTTTGTT TTCTTTTTTC GTTTTCTTTG CAATTTTTTT TGCTAAACGG TAGGTTTCAT GTTCGTTTGT
      Ac3  AAAACTCCAC AAAATGATAC GGCCAGGTGG TAGAAAAGTA GGAAAACTTA GAAAAATAAG GAACAAGAGA TGATTAATAA TCCACGGGAT GAGCTCTTGT
```

```
         301                                                                                                                  400
T26a8    GTTTTTACAG CTATGTGGTC GTGGAATAAC TTCAACTGAC TACGAAAAAG GACAAATTGT GCCAAATTAT CTCAAGGCTT CTCGAATCGT CAGATTTTTC
Zc395    GTTTTTACAG CTATG.GGTC GTGGAATAAC TTCAACTGAC TACGAAAAAG GACAAATTGT GCCAAATTAT CTCAAGGCTT CTCGAATCGT CAGATTTTTC
C33h5    GTTTTTACAG CTATG.GGTC GTGGAATAAC TTCAACTGAC TACGAAAAAG GACAAATTGT GCCAAATTAT CTCAAGGCTT CTCGAATCGT CAGATTTTTC
F48e8    GTTTTTACAG CTATG.GGTC GTGGAATAAC TTCAACTGAC TACGAAAAAG GACAAATTGT GCCAAATTAT CTCAAGGCTT CTCGAATCGT CAGATTTTTC
Tc6.1    GTTTTTACAG CTATG.GGCC GTGGAATAAC TTTAACTGAC AACGAAAAAG GACAAATTGT GCAAAATTAT CTCAAGGCTT CTCGGATCGT CAGATTTTTC
Zk669rc  GTTTTTACAG CTATG.GGCC GTGGAATAAC TTTAACTGAC AACGAAAAAG GACAAATTGT GCAAAATTAT CTCAAGGCTT CTCGGATCGT CAGATTTTTC
Zk180rc  GTTTTTACAG CTATG.GGCC GTGGAATAAC TTTAACTGAC AACGAAAAAG GACAAATTGT GCAAAATTAT CTCAAGGCTT CTCGGATCGT CAGATTTTTC
W03a3    GTTTTTACAG CTATG.GGTC GTGGAATAAC TTTAACTGAC TACGAAAAAG GACAAATTGT GCAAAATTAT CTCAAGGCTT CTCGGATCGT CAGATTTTTC
F53b7    GTTTTTACAG CTATG.TGTC GTGGAATAAC TTTAACTGAC TACGAAAAAA GACGAATTGT GCAAAATTAT CTCAAGGCTT TTCGGATCGT CAGATTTTTC
Ac3      ACTAATCACA TCAAAAACGC GTGAATCCAT GAAAAGTTAA AGATTGTCCA GGAGAGTTTG GTCAACTTTA TGCCATGCGT CAAGAATTG. ..CTCTTTTG

         401                                                                                                                  500
T26a8    GTGATTTGAA ACTTTCGAGA GATATGATCG ATCGATTTC. .......... .......... .......... .......... .......... ..........
Zc395    GTGATTTGAA ACTTTCGAGA GATATGATCA CTCGATAT.. .......... .......... .......... .......... .......... ..........
C33h5    GTGATTTGAA ACTTTCGAGA GATATGATCA CTCGATATGC TTCCAATCCT GCCGCTTATT GCACAAAAAA GTCTTCTGGT CGCACACCTT TTCTTTCTGT
F48e8    GTGATTTGAA ACTTTCGAGA GATATGATCA CTCGATATGC TTCCAATCCT GCCGCTTATT GCACAAAAAA GTCTTCTGGT CGCACACCTT TTCTTTCTGT
Tc6.1    GTGATTTGAA ACGTTTGAGG GATATGATCA CTCGATATGC TTCAAATCCT GCCGCTTATT GCACCAAAAA GTCTTCTGGT CGCCCACCAC TCCTTTCTGG
Zk669rc  GTGATTTGAA ACGTTTGAGG GATATGATCA CTCGATATGC TTCAAATCCT GCCGCTTATT GCACCAAAAA GTCTTCTGGT CGCCCACCAC TCCTTTCTGG
Zk180rc  GTGATTTGAA ACGTTTGAGG GATATGATCA CTCGATATGC TTAAAATCCT GCCGCTTATT GCACCAAAAA GTCTTCTGGT CGCCCACCAC TCCTTTCTGG
W03a3    GTGATTTGAA ACGTTTGAGA GATATGATCA CTCGATATGC TTCAAATCCT GCCGCTTATT .....TTTTA GTCTTCTGGT CGCCCACCAC TCCTTTCTGG
F53b7    GTGATTTGAA ACGTTCGAGA GATATGATCA TTCGATATGC TTCAAATCCT GCCGCTTATT GCACCAAACA GTCTTCTGGT CGCCCACCAC TTCTTTCTGG
Ac3      AGGTCGTTAA CATTGCTATT GTTTCCATCG TAAATCTTTC TCATCATGAA TCCCCAAACG TTTTAAATTA TGTTTATGTC CGGGGAGCAG GCAGGCGAGT

         501                                                                                                                  600
T26a8    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
Zc395    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
C33h5    CAGACCCAAG CGAAAAATCG TTCGTCGAGC ATCCAATTGA ACTTTTTGAC TTGCTCGAAA AGTAGGAGCG AGATGAACCT GCCAGTGTCT GTTGAGGCCG
F48e8    CAGACCCAAA AAAAAAACCG TTCGTCGAGC ATCCAATTGA ACAGT..GAC TTGCTCGAAA AGTAGGAGCG AGATGAACCT GGCAGTGTCT GTTGAGGCCG
Tc6.1    TAGAGACAAG CGAAAAATCG TTCGTCGAGC ATTAAATTGA ACAGT..GAC TTGCTCGAAA AGTAGGAGCG AGATGAACCT GCCAGTGTCT GTTGAGACCG
Zk669rc  TAGAGACAAG CGAAAAATCG TTCGTCGAGC ATTAAATTGA ACAGT..GAC TTGCTCGAAA AGTAGGAGCG AGATGAACCT GCCAGTGTCT GTTGAGACCG
Zk180rc  TAGAGACAAG CGAAAAATCG TTCGTCGAGC ATTAAATTGA ACAGT..GAC TTGCTCGAAA AGTAGGAGCG AGATGAACCT GCCAGTGTCT GTTGAGACCG
W03a3    TAGAGACAAG CGAAAAATCG TTCGTCGAGC ATTCAATTGA ACAGT..GAC TTGCTCGAAA AGTAGGAGCG AGATAAACCT GCCAGTGTCT GTTGAGACCG
F53b7    TAGAGACAAG CGAAAAATCG TTCGTCGAGC ATCAAATTGA ACAGT..GAC TTGCTCGAAA ATTAGGAGCG AGATGCACCT GCCAGTGTCT GTTGAGACTG
Ac3      CGGGAACAGG ....GAGCCC TCTGTCACTT ATCCACTTG. ATTGTACGAC GCGCGCGGTG ACAAGGCGCT CCATCTTGTT G......... ..........
```

```
         601                                                                                                                  700
  T26a8  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  Zc395  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  C33h5  TACGACATGT TCTTTTGAAG TCTCAGTTTA ACAAAAGACG AAAATTAAGA AAGGCTCCAT TCATTACC.. AAAAAACCGC CAAAACCGTA TTCAGTTTGC
  F48e8  TACGACATGT TCTTTTGAAG TCTCAGTTTA ACAAAAGACG AAAATTAAGA AAGGCTCCAT TCATTACC.. AAAAAACCGC CAAAACCGTA TTCAGTTTGC
  Tc6.1  TACGACGTG. TCCTTCGAAG TCCCAGTTTA TCAAAAGACG AAAATTAATA AAGGCTAATT TCATTACCGA AAAACACTGC CAAAATCGTA TTCAGTTTGC
Zk669rc  TACGACGTG. TCCTTCGAAG TCCCAGTTTA TCAAAAGACG AAAATTAATA AAGGCTAATT TCATTACCGA AAAACACTGC CAAAATCGTA TTCAGTTTGC
Zk180rc  TACGACGTG. TCCTTCGAAG TCCCAGTTTA TCAAAAGACG AAAATTAATA AAGGCTAATT TCATTACCGA AAAACACTGC CAAAATCGTA TTCAGTTTGC
  W03a3  TACGTCGTG. TCCTTCGAAG TCCCAGTTTA TCAAAAGACG AAAATTAATA AAGGCTTATT TCATTACCGA AAAACACTGC CAAAATCGTA TTCAGTTTGC
  F53b7  CACGACGTGT TCGTTCGAAT TCCCAGTTTA TCAAAAGACG AAAATTGAGA AAGGCTCCTT TCATTACCGA AAAA.ACCGC TAAAATCGTA TTCAGTTAGC
    Ac3  .......... .......... .......... .......... ........GT AAGTGTATTT TCTGGACCTG TTTCGACGGA TAA....... ..........

         701                                                                                                                  800
  T26a8  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  Zc395  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  C33h5  TAAAATGAGT CAGGGAACTA ACTGAAGACA AGTGAGGATT ACGGTATAAT CATTCAAGCC CAGTTTTTGG TTTCAGTTCA TCTTTT.CTT TTCTCAAAGC
  F48e8  TAAAATCAGT CAGGGAACTA ACTGAAGACA AGTGAGGATT ACGGTATAAT CATTCAAGCC CAGTTTTTGG TTTCAGTTCA TCTTTT.CTT TTCTCAAATC
  Tc6.1  TAAAATCAGC CAGAGAACTA AC.GGAGACA AGTGAGGATT ATGGTATAAT CATTCAAGCC CAGTTTTTGG TTTCAGATCA TCTTTT.CTT TTCTCAAATC
Zk669rc  TAAAATCAGC CAGAGAACTA AC.GGAGACA AGTGAGGATT ATGGTATAAT CATTCAAGCC CAGTTTTTGG TTTCAGATCA TCTTTT.CTT TTCTCAAATC
Zk180rc  TAAAATCAGC CAGAGAACTA AC.GGAGACA AGTGAGGATT ATGGTATAAT CATTCAAGCC CAGTTTTTGG TTTCAGATCA TCTTTT.CTT TTCTCAAATC
  W03a3  TAAAATCAGC CAGAGAACTA AC.TGAGACA AGTGAGGATT ATGGTATAAT CATTCAAGCC CAGCTTTTGG TTTCAGATCA TCCTTT.CTT TTCTCAAATC
  F53b7  TAAAATCAGC CAGAGAACTA ACTGGAGACA AGTGAAGAGT ACCGTATGAT CATTCAAGCC CA.ATTTTGG TTTTAGTTCA TCTTTTCCTT TTCTTAAATC
    Ac3  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........

         801                                                                                                                  900
  T26a8  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  Zc395  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
  C33h5  GTGC..CTAA TCACGGTAGT AATCTGGTTC ATCACAGTTA AACTTTTTCT CGTCACTGAA TGAGAGTATG .......... .......... ..........
  F48e8  GTGCCACTAA TCACGGTAGT AATCTGGTTC ATTACAGTTA AACTTTTTCT CGTCACTGAA GATGAACTGA AACCAAAAAC TGGGCTTGAA TGATTATACC
  Tc6.1  GTGCCAGTAA TCACGGTAGC CATCAGGACC ATCACAGTTA AACTTTTTCT CGCCACTGAA GATGAACTGA AACCAAAAAC TGGGCTTGAA TGATTATACC
Zk669rc  GTGCCAGTAA TCACGGTAGC CATCAGGACC ATCACAGTTA AACTTTTTCT CGCCACTGAA GATGAACTGA AACCAAAAAC TGGGCTTGAA TGATTATACC
Zk180rc  GTGCCAGTAA TCACGGTAGC CATCAGGACC ATCACAGTTA AACTTTTTCT CGCCACTGAA GA.....TGA AACCAAAAAC TGGGCTTGAA TGATTATACC
  W03a3  GTGCCAGTAA TCACGGTAGC CATCAGGACC ATCACAGTTA AACTTTTTCT CGTCACTGAA GATGATCTGA GACCAAAACA TGAGATTGAA TGATCATACC
  F53b7  GTACCAGTAA TCACGGTAGC CTTCTGGTTC ATCACAGTTA AACTTTTCCT CGTCACTGAA GATGATCTGA GACCAATACA TGAGCTTGAA TGATCATACC
    Ac3  .......... ..ACGGTATC AAACCGCGCC TCATCACATA TCTGTAATCG CATC...... .......... .......... .......... ..........
```

```
         901                                                                                                                  1000
T26a8    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
Zc395    .......... .......... .......... .......... .......... .......... .......... .......... .......... ....ATAATT
C33h5    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
F48e8    GTAATCCTCA CTTGTCTTCA GTTAGTTCCC TGACTGATTT TAGCAAACTG AATACGGTTT TGGC..GGTT TTTTGGTAAT GAATGGAGCC TTTCTTAATT
Tc6.1    ATAATCCTCA CTTGTCTCC. GTTAGTTCTC TGGCTGATTT TAGCAAACTG AATACGATTT TGGCAGTGTT TTTCGGTAAT GAAATTAGCC TTTATTAATT
Zk669rc  ATAATCCTCA CTTGTCTCC. GTTAGTTCTC TGGCTGATTT TAGCAAACTG AATACGATTT TGGCAGTGTT TTTCGGTAAT GAAATTAGCC TTTATTAATT
Zk180rc  ATAATCCTCA CTTGTCTCC. GTTAGTTCTC TGGCTGATTT TAGCAAACTG AATACGATTT TGGCAGTGTT TTTCGGTAAT GAAATTAGCC TTTATTAATT
W03a3    GTAATCCTCA CTTGTCTCCA GTTAGTTCTC TGGCTGATTT TAGCAAACTG AATACGATTT TGGCAGTGTT TTTCGGTAAT GAAATAAGCC TTTATTAATT
F53b7    GTAATCCTCA CTTGTCTCCA GTTAGTTCTC TGGCTGATTT TAGCAAACTG AATACGACTT TGGCAGTGTT TTTCGGTAAT GAAAGGAGCC TTTCCTAATT
Ac3      .......... .......... .......... .......... .CGCAAACTG GATACGATTT TGGCGGTGTT TTTTGGTGAC GCATGGAGCC TTTCTCAATT

         1001                                                                                                                 1100
T26a8    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
Zc395    TTCGTCTTTT GTTAAACTGA GACTTCAAAA GAACATGTCG TACGGCCTCA ACAGACACTG CCAGGTTCAT CTCGCTCCTA CTTTTCGAGC AAGTCACTGT
C33h5    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
F48e8    TTCGTCTTTT GTTAAACTGA GACTTCAAAA GGACATGTCG TACGGCCTCA ACAGACACTG CCAGGTTCAT CTCGCTCCTA CTTTTCGAGC AAGTCACTGT
Tc6.1    TTCGTCTTTT GATAAACTGG GACTTCG.AA GGACACGTCG TACGGTCTCA ACAGACACTG GCAGGTTCAT CTCGCTCCTA CTTTTCGAGC AAGTCACTGT
Zk669rc  TTCGTCTTTT GATAAACTGG GACTTCG.AA GGACACGTCG TACGGTCTCA ACAGACACTG GCAGGTTCAT CTCGCTCCTA CTTTTCGAGC AAGTCACTGT
Zk180rc  TTCGTCTTTT GATAAACTGG GACTTCG.AA GGACACGTCG TACGGTCTCA ACAGACACTG GCAGGTTCAT CTCGCTCCTA CTTTTCGAGC AAGTCACTGT
W03a3    TTCGTCTTTT GATAAACTGG GACTTCG.AA GGACACGACG TACGGTCTCA ACAGACACTG GCAGGTTCAT CTCGCTCCTA CTTTTCGAGC AAGTCACTGT
F53b7    TTCGTCTTTT GATAAACTGA GACTTCGAAA GAACACGTCG TGCAGTCTCA ACAGACACTG GCAGGTTCAT CT........ .TTTTTGAGC AAGTCGCTGT
Ac3      TTCTTCTTTT AATAAATTGG GACTTCGAAA GAACTCGTCG TACGGTCTCA ACATACACTG GCAGGTTCAT CTCGCTCTTT ATTTTCGAGC AAGTCACTGT

         1101                                                                                                                 1200
T26a8    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
Zc395    TCAATTGGAT GCTCGACGAA CGGTTTTTTT TTTGGGTCTG ACAGAAAGAA AAGGTGTGCG ACCAGAAGAC TTTTTTGTGC AATAAGCGGC AGGATTGGAA
C33h5    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
F48e8    TCAATTGGAT GCTCGACGAA CGGTTTTTTT TTTGGGTCTG ACAGAAAGAA AAGGTGTGCG ACCAGAAGAC TTTTTTGTGC AATAAGCGGC AGGATTGGAA
Tc6.1    TCAATTTAAT GCTCGACGAA CGATTTTTCG CTTGTCTCTA CCAGAAAGGA GTGGTGGGCG ACCAGAAGAC TTTTTGGTGC AATAAGCGGC AGGATTTGAA
Zk669rc  TCAATTTAAT GCTCGACGAA CGATTTTTCG CTTGTCTCTA CCAGAAAGGA GTGGTGGGCG ACCAGAAGAC TTTTTGGTGC AATAAGCGGC AGGATTTGAA
Zk180rc  TCAATTTAAT GCTCGACGAA CGATTTTTCG CTTGTCTCTA CCAGAAAGGA GTGGTGGGCG ACCAGAAGAC TTTTTGGTGC AATAAGCGGC AGGATTTGAA
W03a3    TCAATTGAAT GCTCGACGAA CGATTTTTCG CTTGTCTCTA CCAGAAAGGA GTGGTGGGCG ACCAGAAGAC T.....AAAA AATAAGCGGC AGGATTTGAA
F53b7    TCAATTGGAT GCTAGACGAA CGATTTTTCG CTTGTCTCTA CCAGAAAGAA GTGGTGGGCG ACCAGAAGAC TTTTT..... AATAAGTGGC AGGATTTGAA
Ac3      TGAATTGGAT GCTCGACAAA CGATATTCCG CTTGTCCCGA TCAGAAAAAA GTGGTGGGCG ACCAGAAGAC TTTTTGGGGC CATAAGCCGC AGGATTGGAA
```

```
         1201                                                                                                                  1300
T26a8    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
Zc395    GCATATCGAG TGATCATATC TCTCGAAAGT TT.CAAATCA CGAAAAATCT GACGATTCGA GAAGCCTTGA GATAATTTGG CACAATTTGT CCTTTTTCGT
C33h5    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
F48e8    GCATATCGAG TGATCATATC TCTCGAAAGT TTCCAAATCA CGAAAAATCT GACGATTCGA GAAGCCTTGA GATAATTTGG CACAATTTGT CCTTTTTCGC
Tc6.1    GCATATCGAG TGATCATATC CCTCAAACGT TT.CAAATCA CGAAAAATCT GACGATCCGA GAAGCCTTGA GATAATTTTG CACAATTTGT CCTTTTTCGT
Zk669rc  GCATATCGAG TGATCATATC CCTCAAACGT TT.CAAATCA CGAAAAATCT GACGATCCGA GAAGCCTTGA GATAATTTTG CACAATTTGT CCTTTTTCGT
Zk180rc  GCATATCGAG TGATCATATC CCTCAAACGT TT.CAAATCA CGAAAAATCT GACGATCCGA GAAGCCTTGA GATAATTTTG CACAATTTGT CCTTTTTCGT
W03a3    GCATATCGAG TGATCATATC TCTCAAACGT TT.CAAATCA CGAAAAATCT GACGATCCGA GAAGCCTTGA GATAATTTTG CACAATTTGT CCTTTTTCGT
F53b7    GCATATCGAG TGATCATATC TCTTGAACGT TT.CAAATCA CGAAAAATCT GACGATCCGA AAAGCCTTGA GATAATTTTG CACAATTCGT CTTTTTTCGT
Ac3      GCATATCGAG AAATCATATC TCTCGAACGA TT.CAAATTA CGAGAAATCT GACGATGCGA TGTGGAGCAC .......... .......... ..........

         1301                                                                                                                  1400
T26a8    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
Zc395    AGTCAGTTGA AGTTATTCCA CGACCCATAG CTGTAAAAAC ACAAACGAAC ATGAAACTTA ACGTTTAGCA AAAAAAATTG CAAAGAAAAC GAAAAAAGAA
C33h5    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
F48e8    AGTCAGTTGA AGTTATTCCA CGACCCATAA CTGTAAAAAC ACAAACGAAC ATGAAACTTA ACGTTTAGCA AAAAAAATTG CAAAGAAAAC GAAAAAAGAA
Tc6.1    TGTCAGTTAA AGTTATTCCA CGGCCCATAG CTGTAAAAAC ACAAACGAAC ATGAAACCTA ACGTTTAGCA AAAAAAATTG CAAAAAAAAC GAAAAAAGAA
Zk669rc  TGTCAGTTAA AGTTATTCCA CGGCCCATAG CTGTAAAAAC ACAAACGAAC ATGAAACCTA ACGTTTAGCA AAAAAAATTG CAAAAAAAAC GAAAAAAGAA
Zk180rc  TGTCAGTTAA AGTTATTCCA CGGCCCATAG CTGTAAAAAC ACAAACGAAC ATGAAACCTA ACGTTTAGCA AAAAAAATTG C.AAAAAAAC GAAAAAAGAA
W03a3    AGTCAGTTAA AGTTATTCCA CGACCCATAG CTGTAAAAAC ACAAACGAAC ATGAAACCTA ACGTTTAGCA AAAAAAATTG CAAAAAAAAC GAAAAAAGAA
F53b7    AGTCAGTTAA AGTTATTTCA CGACACATAG CTGTAAAAAC ACAAACGAAC ATGAAACCTA ACGTTTAGCA AAAAAAATTG CAAAGAAAAC GAAAAAAGAA
Ac3      .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........

         1401                                                                                                                  1500
T26a8    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
Zc395    AACGAAAACC AAGGGTGGCC GTATCATTTT GTGGATTTTT T.AATATGGT TTTTCATGAA AAAAATGGTT TGTCTTTTTT CTGAACGATT CAGGTATATT
C33h5    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
F48e8    AACGAAAACC AAGGGTGGCC GTATCATTTT GTGGATTTTT TAAATATGGT TTTTCATGAA AAAATGGTTT ATC.TTTTTT ACAAACGGTT CAGGTATATT
Tc6.1    AACAAAAACC AAGGGTGGCC GTATCATTTT GTGGATTTTT T.TATATGGT TTTTCATGAA AAAATGGTTT ATCTTTTTTT ACGAACGGTT CAGGTATATT
Zk669rc  AACAAAAACC AAGGGTGGCC GTATCATTTT GTGGATTTTT T.TATATGGT TTTTTATGAA AAAATGGTTT ATCTTTTTTT ACGAACGGTT CAGGTATATT
Zk180rc  AACAAAAACC AAGGGTGGCC GTATCATTTT GTGGATTTTT T.TATATGGT TTTTCATGAA AAAATGGTTT ATCTTTTTTT ACGAACGGTT CAGGTATATT
W03a3    AACAAAAACC AAGGGTGGCC GTATCATTTT GTGGA.TTTT T.TATATGGT TTTTCATGAA AAAATGGTTT ATCTTTTTTT ACGAACGGTT CAGGTATATT
F53b7    AACAATAACC AAGGGTGGCC GTATCATTTT GTGGATTTTT T.TATATGGA TTTTCATGAA AAAATGGTTT ATCTTTTTTT ACGAACGGTT CATGTATATT
Ac3      .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
```

```
         1501                                                                                                                  1600
T26a8    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
Zc395    TCCAATACTT ATCTGAAAGT TAGCCAGTAC AGCGAACATT TTCACAAGAA ACTATGTTTT GCTATCTCAA CCCAGTTTTG AGTTATACCA AAATTTGGGG
C33h5    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
F48e8    TCCAATACTT ATCTGAAAGT AAGTTAGTAC AGCGAACATT TTCACAAGAA ACTATGTTTT GCTATCTCAA CCCAGTTTTG AGTTATACCA AAATTT.GGG
Tc6.1    TCCAATACTT ATCTGAAAGT TAGCCAGTAC AGCGAACATT TTCACAAGAA ACTATGTTTT GCTATCTCAA CCCAGTTTTG AGTTATACCA AAATTTGGGG
Zk669rc  TCCAATACTT ATCTGAAAGT TAGCCAGTAC AGCGAACATT TTCACAAGAA ACTATGTTTT GCTATCTCAA CCCAGTTTTG AGTTATACCA AAATTTGGGG
Zk180rc  TCCAATACTT ATCTGAAAGT TAGCCAGTAC AGCGAACATT TTCACAAGAA ACTATGTTGT GCTATCTCAA CCCAGTTTTG AGTTATACCA AAATTTGGGG
W03a3    TCCAATACTT ATCTGAAAGT TAGCCAGTAC AGCGAACATT TTCACAAGAA ACTATGTTTT GCTATCTCAA CCCAGTTTTG AGTTATACCA AAATTTGGGG
F53b7    TCCAATACTT ATCTGAAAGT TAGCCAGTAC AGCGGACATT TTCACAAGAA ATTATGTTTT GCTATCTCAA CCCAGTTTTG AGTTATACCA AAATTTGGGG
Ac3      .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........

         1601                                1637
T26a8    .......... .......... .......... .......
Zc395    GTGGCCGTAT CATTATGTGG AGCACCTCGG TAAAA..
C33h5    .......... .......... .......... .......
F48e8    GTGGCCGTAT CATTATGTGG AGCACTGTAG TATCGAT
Tc6.1    GTGGCCGTAT CATTATGTGG AGCACTG... .......
Zk669rc  GTGGCCGTAT CATTATGTGG AGCACTGTAG TAACTTA
Zk180rc  GTGGCCGTAT CATTATGTGG AGCACTGTAT ATGTATA
W03a3    GTGGCCGTAT CATTATGTGG AGCACTGTAC TAAGAAA
F53b7    GTGGCCGTAT CATTATGTGG AGCACTGTAT AAAATGA
Ac3      .......... .......... .......... .......
```