

Winter 2006

Transposable elements: What have you done for me lately? A genomics based investigation into the potential functional roles of transposable elements using the model organism *Caenorhabditis elegans*

Sarah Prescott Kenick
University of New Hampshire, Durham

Follow this and additional works at: <https://scholars.unh.edu/dissertation>

Recommended Citation

Kenick, Sarah Prescott, "Transposable elements: What have you done for me lately? A genomics based investigation into the potential functional roles of transposable elements using the model organism *Caenorhabditis elegans*" (2006). *Doctoral Dissertations*. 353.
<https://scholars.unh.edu/dissertation/353>

This Dissertation is brought to you for free and open access by the Student Scholarship at University of New Hampshire Scholars' Repository. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of University of New Hampshire Scholars' Repository. For more information, please contact nicole.hentz@unh.edu.

TRANSPOSABLE ELEMENTS: WHAT HAVE YOU DONE FOR ME LATELY?
A GENOMICS BASED INVESTIGATION INTO THE POTENTIAL FUNCTIONAL
ROLES OF TRANSPOSABLE ELEMENTS USING THE MODEL ORGANISM
CAENORHABDITIS ELEGANS

BY

Sarah Prescott Kenick

B.S. Worcester Polytechnic Institute, 1993

M.Ed. University of New Hampshire, 1999

DISSERTATION

Submitted to the University of New Hampshire

In Partial Fulfillment of

The Requirements for the Degree of

Doctor of Philosophy

In

Genetics

December 2006

UMI Number: 3241644

INFORMATION TO USERS

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 bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send 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.

UMI[®]

UMI Microform 3241644

Copyright 2007 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

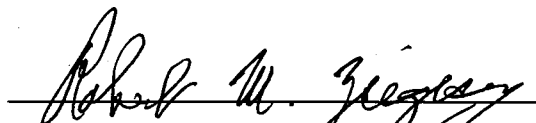
This dissertation has been examined and approved.



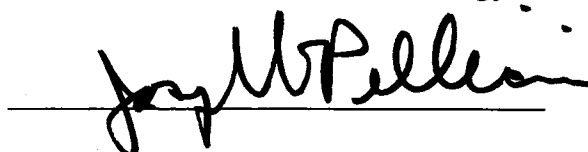
Dissertation Director, John J. Collins
Associate Professor of Biochemistry
and Molecular Biology



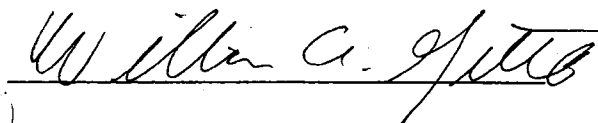
Andrew P. Laudano
Associate Professor of Biochemistry
and Molecular Biology



Robert M. Zsigray
Professor Emeritus of Microbiology



Joseph G. Pelliccia
Associate Professor of Biology



William Gilbert
Guest Professor, Hubbard Center for
Genome Studies

9/15/06

Date

DEDICATION

To Joseph and Alexandra, for the motivation they never knew they gave.

ACKNOWLEDGEMENTS

I wish to extend my most sincere thanks to many people who have provided support and guidance throughout this project. I would like to thank my advisor, John Collins, for introducing me to *C.elegans*, and for guiding me throughout my graduate work here at UNH. I would also like to thank other members of my committee: Andy Laudano, Joe Pelliccia, Bob Zsigray, and Will Gilbert, for their support.

I also wish to thank other members of the UNH community who have helped along the way. Thanks to Harry Richards, Cari Moorhead, Sharon Andrews, and the entire UNH Graduate School office for providing financial and personal support throughout this project.

I also wish to thank Sylvia Fischer for providing me with supplemental information regarding Tc's 9 and 10, as well as assisting with troubleshooting some of the data mining parts of this project. I wish to thank Lincoln Stein and Todd Harris for their assistance with my annotation and mining of WormBase, and for their permission to utilize graphics in WormBase in this dissertation.

I could not have been successful in this work without the personal friendship and support of Ryan Sweeder, Stephen Druschel, and Holly Tomilson, each of whom provided much needed motivation along the way.

Lastly, none of this work would have been possible without the never-ending love and support from my family. Thanks to my children, Joseph and Alexandra, to whom this thesis is dedicated, for their understanding of the long hours necessary to complete this work. Thanks to my sis April, who always was there to listen. Thanks to my father, for his support and strength. Finally, a very big thank you to my husband, Joe, who always believed in my ability to succeed.

TABLE OF CONTENTS

DEDICATION	iii
ACKNOWLEDGEMENTS.....	iv
TABLE OF CONTENTS.....	vi
LIST OF FIGURES.....	ix
LIST OF TABLES	xi
ABSTRACT.....	xii
CHAPTER.....	PAGE
I. INTRODUCTION.....	1
General Background	1
<i>Class II Elements</i>	2
<i>The Tc1 and Tc3 transposons</i>	3
<i>Other active transposons</i>	4
<i>Transposons with no detected activity</i>	6
Mechanisms of Class II transposition.....	7
<i>Transposon insertion target</i>	8
Genome Architecture of <i>C.elegans</i>	9
<i>Gene density and evolution</i>	10
<i>Gene organization</i>	13
Comparative Genomics – <i>Caenorhabditis briggsae</i> (<i>C.briggsae</i>)	16

Regulation of transposition.....	19
RNAi	20
Heterochromatin silencing.....	24
II. METHODS	27
Bioinformatics – Locations and annotation of elements	27
Sequence Analysis	30
III. RESULTS.....	33
Global overview of locations of transposable elements in <i>C.elegans</i>	35
<i>Location by TE family</i>	35
<i>Distribution of Fragment elements</i>	37
<i>Location within each Linkage Group</i>	41
<i>Genetic position (intergenic, in introns, exons. etc.) of elements</i>	47
<i>Operons and transposable elements</i>	48
Comparative Genomics – <i>C.elegans</i> and <i>C.briggsae</i>	49
<i>Transposable element presence</i>	50
Intron Studies.....	51
<i>C.elegans Intron Study (E Intron Study)</i>	52
<i>C.briggsae Intron Study (B Intron Study)</i>	66
Additional Analysis.....	74
Future Directions.....	74
IV. DISCUSSION.....	76
REFERENCES	85
APPENDICES	100

APPENDIX A: LINKAGE GROUP MAPS.....	101
APPENDIX B: INTRON STUDIES	118
<i>C.elegans</i> Intron Study.....	119
<i>C.briggsae</i> Intron Study.....	157

LIST OF FIGURES

Figure 1 Active Transposon Families.	4
Figure 2 Mechanism of Tc3 Transposition.....	8
Figure 3 RNAi mechanisms in <i>C.elegans</i>	23
Figure 4 Example of IRs with GAP	39
Figure 5 Example of FRAG	40
Figure 6 Screenshot of Tc 4, 4v, 5, 9.....	41
Figure 7 Overall position of all TE's.....	43
Figure 8 Locations of Active Transposable Elements (Fischer et. al. 2003)	44
Figure 9 Example of a conserved intron (C31A11.7/CBG24402)	55
Figure 10 T05H4.10 Part I –element created intron example.....	62
Figure 11 T05H4.10 Part II – element created intron example.....	63
Figure 12 T05H4.10 Part III – element created intron example.....	63
Figure 13 clec-41 Part I – element created intron example	64
Figure 14 clec-41 Part II – element created intron example	65
Figure 15 clec-41 Part III – element created intron example	65
Figure 16 CBG 07789 – Conserved intron example	67
Figure 17 CBG07789/F02C12.1 pair – Conserved intron example	68
Figure 18 CBG09294 - Conserved intron?	68
Figure 19 CBG05090 - Conserved intron?	71
Figure 20 CBG06979 - Element created intron.....	72

Figure 21 CBG06979 - Element created intron.....	72
Figure 22 CBG06979 - Element created intron.....	73

LIST OF TABLES

Table 1 Chromosomal distributions of protein-coding genes ¹	11
Table 2 Abbreviations used	29
Table 3 Tc element sequence source list	29
Table 4 Locations of all Transposable Elements	36
Table 5 Distribution of element fragments	38
Table 6 Calculated Gene Density	45
Table 7 Table 1 repeated ¹	46
Table 8 Tabulation of genetic positions of TE's	47
Table 9 <i>C.briggsae</i> full elements	51
Table 10 <i>C.elegans</i> Intron Study Part I	57
Table 11 <i>C.elegans</i> Intron Study Part II	58
Table 12 <i>C.briggsae</i> Intron Study	69

ABSTRACT

TRANSPOSABLE ELEMENTS: WHAT HAVE YOU DONE FOR ME LATELY?
A GENOMICS BASED INVESTIGATION INTO THE POTENTIAL FUNCTIONAL
ROLES OF TRANSPOSABLE ELEMENTS USING THE MODEL ORGANISM
CAENORHABDITIS ELEGANS

by

Sarah Prescott Kenick

University of New Hampshire, December, 2006

The genomes of all organisms contain discrete DNA sequences present as dispersed repetitive elements called transposons. Transposons have the unique ability to move to new chromosomal locations. Problems of uncontrolled movement of transposons can result in mutations, rearrangement, and even broken chromosomes. Often termed "selfish parasites" that invade a host genome, there is a longstanding question of whether they have a functional role. As a first step in an effort to investigate this question, I identified and annotated 276 full length and partial elements in the *C.elegans* genome. I determined the genomic location of each and looked for patterns resulting from their presence. I found that they are widespread throughout the *C.elegans* genome, and do not cluster on the arms of the chromosomes as was previously thought. In addition, I have found examples of elements that have created introns in *C.elegans* genes and for which there are conserved introns in a closely related species,

C. briggsae. Lastly, I have discovered evidence of potential novel intron creation by transposable elements in both *C. elegans* and *C. briggsae*. These results establish evidence for the genome's adaptation to the presence of these elements, and point to the possibility of the host genome utilizing their unique characteristics to regulate gene expression.

CHAPTER I

INTRODUCTION

General Background

The genomes of all organisms contain discrete DNA sequences present as dispersed repetitive elements called transposons. These elements have the unique ability to move to new chromosomal locations. Movement of transposons can result in mutations, rearrangement, and even broken chromosomes. Thus, regulating the activity of transposons is important for maintaining genome integrity. Understanding the role of transposable elements in the host genome is the focus of my thesis.

Approximately 12 % of the *C.elegans* genome is derived from transposable elements (*C.elegans* Sequencing Consortium 1998; Sijen and Plasterk 2003; Stein et al., 2003). Transposons are broadly classified into two classes according to their general structure and mode of transposition (reviewed in Finnegan 1989; Berg and Howe 1989).

Class I Elements

Class I elements are commonly referred to as retrotransposons because they resemble retroviruses in their structure and mode of transposition (Boeke et al.,

1985; Garfinkel et al., 1985, Berg 1989). They encode element-specific proteins including a reverse transcriptase (RT) important for transposition. The RT of Class I elements facilitates transposition via a RNA intermediate. These elements are often, but not always, flanked by long terminal repeats (LTRs). Examples of LTR-bound Class I elements include the BS1 elements in maize, copia-like elements in *Drosophila*, Ty elements of *S. cerevisiae*, and THE element in humans. Non-LTR Class I elements include Cin4 of maize and LINEs (long interspersed elements i.e. L1) and SINEs (short interspersed elements i.e. Alu) in humans.

Class II Elements

Class II elements are sequences of variable size characterized by presence of terminal inverted repeats (TIRs). Class II transposable elements are grouped into families according to their ability to interact with each other genetically (Fedoroff 1989). Examples of Class II elements include members of the Tc1/mariner superfamily, as well as *Ac/Ds* (*Activator/Dissociation*) transposon pair in maize, and P elements in *Drosophila*. For many Class II elements, the internal sequences encode a protein involved in element mobility, termed transposase. The subject of my research has been on these class II transposable elements, classified by their characteristic of moving as discrete pieces of DNA. Figure 1 displays the general structure of each of the active element families in the *C.elegans* genome (Tc 1-7), with the shaded in boxes

representing the TIR regions, and the arrows representing the gene encoding the transposase.

The Tc1 and Tc3 transposons

Tc1 and Tc3 are the most active and best-characterized transposons in *C.elegans*. Tc1 was isolated as a repeated sequence responsible for polymorphism among different strains (Emmons et al., 1983; Liao et al., 1983; Rosenzweig et al., 1983). Analysis of spontaneous and reversible mutations of the *unc-54* muscle gene demonstrated the mobile nature of Tc1 (Eide and Anderson 1985; Eide and Anderson 1988). This feature was used to clone another muscle gene, *unc-22*, by transposon tagging (Moerman et al. 1986; Moerman and Waterston 1984). The subsequent characterization of additional spontaneous *unc-22* mutations lead to the identification of Tc3 (Collins et al., 1989). Both Tc1 and Tc3 are found as multiple full length copies dispersed throughout the worm genome. Each member of a family is unique at the sequence level due to single nucleotide polymorphisms.

Tc1 is 1,610 bp long and contains two 54-bp terminal inverted repeats (Rosenzweig et al., 1983). Tc3 is an element of 2,335 bp with 462 bp TIRs. The genome of the Bristol N2 strain contains 31 and 22 copies of Tc1 and Tc3, respectively (Fischer et al., 2003). These numbers are strain dependent. In some strains such as Bergerac, Tc1 transposition is active in the germ line and each haploid genome contains up to 300-500 Tc1 copies (Egilmez et al. 1995; Emmons et al., 1983; Liao et al., 1983).

Tc1 and Tc3 are part of a superfamily of transposable elements, which is named after its two best-studied members: Tc1 and the related transposon *mariner*. Tc1/*mariner* elements are probably the most widespread DNA transposons and can be found in fungi, plants, ciliates, and animals including vertebrates (for review see Plasterk et al., 1999). These transposons are about 1,300-2,400 bp in length, are flanked at either end by TIRs and contain a single open reading frame that encodes a transposase enzyme.

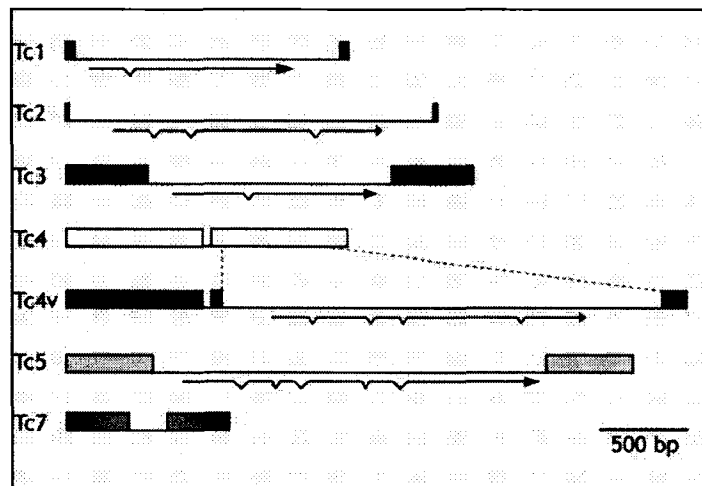


Figure 1 Active Transposon Families. Shaded boxes represent TIR regions, and arrows represent gene encoding transposase protein. Tc7 utilizes the Tc1 transposase for transposition. Adapted from Fischer et. al. (2003) with permission.

Other active transposons

The Tc2 transposon is 2,074 bp in length and has perfect terminal inverted repeats of 24 bp (Ruvolo et al., 1992). Gene prediction algorithms suggest that Tc2 encodes a 477 aa protein containing a DNA binding domain and a catalytic

domain related to the DDE endonuclease superfamily. Transposition of Tc2 has been documented in the offspring of crosses between Bristol N2 and Bergerac BO or in a mut-2 background (Francis et al., 1995; Levitt and Emmons 1989).

Tc4 is a fold-back element of 1.6 kb, which contains almost perfect terminal inverted repeats of 774 bp with a 57-bp unique internal sequence. No open reading frame can be detected within Tc4. A variant class of Tc4 (Tc4v, 5 copies in the N2 genome) contains a 2,343 bp sequence which replaces 477 bp in one of the inverted repeats (See Figure 1) (Li and Shaw 1993). A transcript from Tc4v has been detected. It may encode a 537-aa protein, which resembles transposases of the DDE superfamily. Tc4v might provide in trans the transposase required to mobilize all Tc4 elements. These elements duplicate a 3-bp target sequence TNA upon integration and are mobile in mut-2 (Yuan et al., 1991) and mut-7 (Ketting et al., 1999) mutator backgrounds.

The Tc5 element is present in four copies per haploid genome in the Bristol N2 strain (Collins and Anderson 1994). It is 3,171 bp long and has 491 bp long terminal inverted repeats. Tc5 and Tc4v share common features. Tc5 encodes a putative 532 amino acid transposase, which is overall 33 % identical to the Tc4v transposase. Tc4 and Tc5 TIRs share a few short nucleotide sequences, and integration of Tc5 causes duplication of the same TNA trinucleotide sequence. Tc5 elements are mobile only in mut-2 (Collins and Anderson 1994) and mut-7 (Ketting et al., 1999) backgrounds.

Tc7 is a 921 bp element that uses the Tc1 transposase for transposition (Rezsohazy et al., 1997). It is made up of two 345 bp inverted repeats separated by a unique sequence that does not contain an open reading frame. Thirty-six of the 38 outer base pairs of Tc7 are identical to those of Tc1. Forced expression of Tc1 transposase in somatic cells causes transposition of Tc7 (Rezsohazy et al., 1997). Furthermore, Tc7 is mobile in the germ line in the same backgrounds as Tc1 such as mut-6 and mut-7 lines.

Transposons with no detected activity

The genome of *C.elegans* contains several class II transposons that are not mobile under laboratory conditions. Tc6 (1602 bp) is a fold-back element (Dreyfus and Emmons 1991; Dreyfus and Gelfand 1999). Tc8 is related to the plant Tourist transposon (Le et al., 2001), but I could not locate a reliable source for its sequence, and thus did not analyze Tc8 in this work. The elements Tc9 and Tc10 (two Tc10 elements were identified by Fischer et.al. (2003) and I have termed them Tc10a and Tc10b to distinguish them) were previously identified by genomic analysis using BLAST searches, and are both thought to be related to Tc4v. The lengths of Tc9, 10a, and 10b are 4295, 3546, and 4184 bp respectively. In addition, for both Tc's 9 and 10, fifteen copies of a smaller 1.6kb transposon were found which have TIRs nearly identical to 9 and 10, but which do not encode a transposase (Fischer et al., 2003). Some can bind to any part of

the DNA molecule, and the target site can therefore be anywhere, while others bind to specific sequences.

Mechanisms of Class II transposition

Class II transposons move via a "cut-and-paste" mechanism: transposase binds the TIRs, catalyses excision and subsequent reinsertion into target DNA in a TA dinucleotide, and leaves behind a double-strand DNA break. The DNA break is subsequently repaired by the cellular machinery. The Tc1 transposase is the only factor required in trans to mediate Tc1 transposition (Vos et al., 1993 and 1996). Similar evidence has been obtained for the Tc3 transposase (van Luenen et al., 1993; van Luenen et al., 1994). The Tc1 and Tc3 transposases are 343 and 329 amino acids long, respectively.

Terminal inverted repeats are both necessary and sufficient (in vitro and in vivo) for transposition as long as transposase is provided in trans. Within the TIRs, the first four bases of the transposon and the transposase binding sites located immediately downstream are strictly required for excision (van Luenen et al., 1994; Vos and Plasterk 1994). Transposon excision results from a pair of double-strand breaks at the ends of the inverted repeat. Transposase makes a staggered cut at the target site producing sticky ends, cuts out the transposon and ligates it into the target site. A DNA polymerase fills in the resulting gaps from the sticky ends and DNA ligase closes the sugar-phosphate backbone. Repair of the resulting single-strand gaps causes a duplication of the TA dinucleotide at each transposon end (See Figure 2).

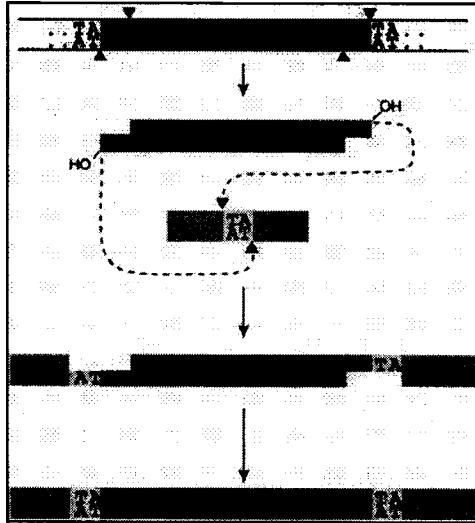


Figure 2 Mechanism of Tc3 Transposition.

Tc3, a member of the Tc1/mariner transposon superfamily, is mobilized by a "cut-and-paste" mechanism. The transposase excises the transposon by causing double-strand breaks at the end of the transposon (arrowheads). The DNA cut is staggered, resulting in a two-base pair 3'-hydroxyl overhang at each terminus. Following excision, transposon then integrates 5' of a thymidine nucleotide at a TA target sequence using the free 3' hydroxyl as a nucleophile. Repair of the resulting single-strand gaps causes a duplication of the TA dinucleotide at each transposon end. Adapted from van Luenen et al., 1994 with permission.

Transposon insertion target

Tc1, Tc3, and Tc7 always integrate into the TA sequence and Tc4 and Tc5 integrate into TNA sites. Since intron sequences are AT-rich, this may explain why such elements have a higher probability of inserting into introns rather than into coding sequences (Martin et al., 2002). Not all TA dinucleotides represent equivalent targets, however. The comparison of Tc1 and Tc3 insertion sites reveals a weak consensus limited to four nucleotides on each side suggesting that the transposase interacts directly with the TA dinucleotide and less specifically with surrounding bases. There also appear to be regional differences in insertion preferences. For example, the gene *unc-22* is hit about a 100 times more frequently than *unc-54* although it's coding sequence is only 3.5 times

larger (Eide and Anderson 1985; Moerman and Waterston 1984). Part of these differences might arise from the fact that transposons such as Tc1 have a preference for local reinsertion into the same chromosome from which they were excised (Fischer et al., 2003). Previous analysis of random insertions indicates the presence of a 4 kb hot spot at the right end of chromosome I, which cannot be explained by local transposition (Granger et al., 2004). My results described in this work do not support this presence of a hot spot (i.e. I did not see a clustering of elements in this area).

Lastly, but most intriguing for my research interests, transposon sequences are not evenly distributed in the genome. For example, previous reports have found them to be located predominantly on the chromosomal arms and in gene poor regions (Fischer et al., 2003). Additionally, a correlation has been found between the density of DNA transposons and the regions of higher chromosomal recombination rate (Duret et al., 2000; Rizzon et al., 2003). These reports led me to looking more closely at exactly where each full and partial fixed transposable element was located. Identifying the locations of all elements should provide further insight into how transposition is regulated.

Genome Architecture of *C.elegans*

As stated above, one of my objectives was to test the model of transposable elements performing a functional role in the genome. A variety of questions came out of the results that I pursued, some of which addressed particular

questions regarding worm genome architecture in the context of transposable elements. For an understanding of these questions and subsequent results, it is necessary to discuss a few key points with regard to the genome of the worm.

The *C.elegans* genome is 100Mb in length, organized into five autosomes (termed Linkage Groups I-V) and one X chromosome. Protein-coding genes are found equally on either strand of DNA and are uniformly distributed throughout the genome. There were 22,227 protein-coding genes found in the Sept 24, 2004 WormBase data release (WS133). They are slightly denser on autosomes than on chromosome X (see Table 1) and, in general, the central regions of the autosomes are denser than the arms. The left arm of chromosome II is an exception.

Gene density and evolution

In more detail, each of *Caenorhabditis C.elegans*' chromosomes is divided into a repeat-poor "central cluster" that rarely undergoes meiotic exchange, and two repeat-rich "arms" that have a ~7-fold higher recombination rate (Barnes et al., 1995; *C.elegans* Sequencing Consortium, 1998). The arms are evolving far

Table 1 Chromosomal distributions of protein-coding genes¹

Chromosome	Size (Mb)	Protein-coding genes	Density (genes/Mb)
I	15.08	3260	216
Left	4.00	685	171
Center	6.26	1573	251
Right	4.82	1002	202
II	15.28	3874	253
Left	5.90	1648	279
Center	5.44	1435	263
Right	3.94	791	201
III	13.76	3103	225
Left	4.80	972	202
Center	4.29	1199	279
Right	4.68	932	199
IV	17.49	3606	206
Left	6.74	1339	198
Center	5.08	1321	260
Right	5.67	946	167
V	20.92	5256	251
Left	6.51	1615	248
Center	6.99	1880	269
Right	7.42	1761	237
X	17.72	3186	180
¹ Adapted from Spieth et. al. (2006) with permission			

more rapidly than the centers of chromosomes, in terms of both substitutions and chromosomal rearrangements such as translocations, inversions, and duplications (*C.elegans* Sequencing Consortium, 1998; Stein et al., 2003). This may be due to a lower tolerance to mutation in the central clusters, which contain most of the essential genes and operons (Blumenthal et al., 2002; Kamath et al., 2003). Alternatively, the arms may simply have a higher mutation rate, since the high recombination rate may provoke substitutions (Cutter and Payseur, 2003), while the abundance of repeats probably triggers chromosomal rearrangements (Coghlan and Wolfe, 2002).

Barnes et al. (1995) noticed that the recombination rate in most *C.elegans* autosomes differs by a factor of ~7–12 between the arms and central clusters. However, in chromosome V, the recombination rate differs by a factor of just four between the arms and cluster. The relatively higher recombination rate in the central cluster of chromosome V may be a cause (or possibly a result) of its "arm-like" characteristics: its high density of gene families (*C.elegans* Sequencing Consortium, 1998), low number of essential genes (Kamath et al., 2003), scarcity of operons (Blumenthal et al., 2002) and abundant species-specific genes (Parkinson et al., 2004).

Gene organization

C. elegans genes in general do not overlap one another, that is to say, their exons do not overlap, but there are numerous examples of either genes that fall within introns of another gene, on the same or the opposite strand. Most *C. elegans* genes are relatively small, covering a genomic region of approximately 3 kb (from start to stop codon including introns); however, there are some very large genes, which skew the average. The median size is only 1,956 bases with a range from 48 to 80,957 bases (genes Y10G3AL.6 and W06H8.8g, respectively) (*C. elegans* Sequencing Consortium, 1998).

There are 126,477 predicted unique, coding exons in the WS133 protein-coding gene set, which account for 25.55% of the genome (*C. elegans* Sequencing Consortium, 1998). The average gene contains 6.4 coding exons; however, there are a few genes with a large number of exons, such as W06H8.8g mentioned above with 66 coding exons. There are also a few single exon genes (570 in WS133) amounting to about 3% of total genes. Almost 60% of these are supported by EST or mRNA data. The average size of unique exons in all protein-coding genes is 208 bases, but there are a small number of very large exons. Again, as with gene size, these few large exons skew the average. The median size is only 123 bases, thus exons are similar in size to exons in human and fly genes (The International Human Genome Sequencing Consortium, 2001).

There are 106,909 predicted unique introns in all of the protein-coding genes of *C.elegans* (WS133 release). Some of these are probably not real introns or have incorrect boundaries because they are either predicted only by Genelocater or based on imperfect alignments of cDNA or single-pass EST reads. Of these, 824 are less than 30 bases, almost all of which probably result from erroneous EST alignments in WormBase. 67,833 introns are considered confirmed because there is EST or cDNA sequences spanning the intron boundaries. The most common size of confirmed introns is 47 bases with the median size being 65 bases. The range of intron size varies from 10 – 21,230 bases (found in mag-1 and kin-1 genes respectively) Intron size in *C.elegans* appears to be positively correlated with local recombination rates (Prachumwat et al., 2004) and short introns are preferentially found in highly expressed genes (Castillo-Davis et al., 2002).

Since a part of my results directed me to look further into the introns of *C.elegans* (and *C.briggsae*), it seems prudent to provide a few more details on the same. The introns of *C.elegans* have always been considered small, but as more genomes are being sequenced and annotated it is becoming evident that they are not distinctly smaller than those of most eukaryotes. The most common size for fly introns is only 59 bases (The International Human Genome Sequencing Consortium, 2001), as compared to 47 bases for the worm. The average size of introns on the largest, macronuclear chromosome of Paramecium is only 25 bases (Zagulski et al., 2004). Fungal introns are also small; Neurospora introns

average 134 bases (Galagan et al., 2003) and *S. macrospora* 106 bases (Nowrousian et al., 2004). In humans, the most common intron size is only 87 bases, but there are also some very large introns, shifting the mean size to more than 3,300 bases (The International Human Genome Sequencing Consortium, 2001).

C.elegans introns follow the GU-AG splice site rule, although GC is a rare 5' splice site variant (Blumenthal and Steward, 1997). From their analysis of 669 introns, Blumenthal and Steward found that *C.elegans* has a highly conserved and extended 3' splice site (UUUCAG) and no obvious polypyrimidine tract other than this 3' splice site consensus. In addition to splicing information, some *C.elegans* introns contain sequences involved in the regulation of gene expression. An example of this is the *pal-1* gene in *C.elegans*, which has a regulatory element in its fifth intron that is responsible for neurogenesis in the male tail of the worm (Zhang and Emmons, 2000).

An unusual and interesting feature of the worm genome is the existence of genes organized into operons. These polycistronic gene clusters contain two or more closely spaced genes, which are oriented in a head to tail direction. They are transcribed as a single polycistronic mRNA and separated into individual mRNAs by the process of trans-splicing (Spieth et al., 1993). There are ~1000 operons in the *C.elegans* genome, of which 96% are conserved in *C.briggsae*, far more than expected if selection did not act to preserve them (60%; Stein et al., 2003).

Gene order in ~15% of the genome is stabilized by selection against rearrangements of operons, since 15% of *C.elegans* genes are part of operons (Blumenthal et al., 2002). In fact, operons are concentrated in the central clusters of *C.elegans* chromosomes, so probably contribute to the lower rearrangement rate in the centers compared to the arms (Blumenthal et al., 2002).

Comparative Genomics – *Caenorhabditis briggsae* (*C.briggsae*)

Another aspect of this project I have pursued has been to compare transposable elements in *C.elegans* to a closely related nematode species, *C.briggsae*. It would thus be important to discuss a few of the similarities/differences between these two species. The *C.briggsae* genome is slightly larger than the *C.elegans* genome (102 vs. 98 Mb), due to a larger amount of repetitive DNA (Stein et al., 2003). They both are predicted to have approximately the same number of genes (19,500 based on the WormBase 2003 estimate). When Stein et al. (2003) compared the genome of *C.elegans* to that of *C.briggsae*; they identified ~4800 conserved segments, with an average size of 37 kb. They estimated that there have been 3.6 interchromosomal rearrangements per Mb in the *C.briggsae* genome (Stein et al., 2003). Thus, an average *C.briggsae* chromosome of ~10-20 Mb consists of a mosaic of ~35-70 segments that correspond to segments from several *C.elegans* chromosomes. A genetic map for *C.briggsae* is currently underway, but at this point is lacking a full assembly of genes in complete linkage groups.

Since *C.elegans* and *C.briggsae* diverged, their chromosomes have been splintered by ~250 reciprocal translocations, ~1400 inversions and ~2700 transpositions (Stein et al., 2003). Intrachromosomal rearrangement is about four times more frequent than interchromosomal rearrangement. Even so, translocations are surprisingly common in *Caenorhabditis* compared to flies, in which translocations are extremely rare (Ranz et al., 2001; Sharakhov et al., 2002). This may be because almost all dipterans have monocentric chromosomes, in which the kinetochores assemble on a localized region in each chromosome. In contrast, species such as *C.elegans* and *C.briggsae* have holocentric chromosomes, where diffuse kinetochores form along the length of each chromosome during mitosis. Since the kinetochores are the primary chromosomal attachment site for spindle microtubules, they play a key role in ensuring high fidelity chromosome transmission in both monocentric and holocentric species. However, little is known of the relationship between the distribution of kinetic activity along chromosomes and the pattern of chromosomal rearrangement.

C.elegans has ~1000 genes not found in *C.briggsae*, and that lack any match in sequence databases (Stein et al., 2003). Of these, ~200 have been confirmed by EST or cDNA data, so are not gene prediction errors. These genes may have diverged so rapidly that their *C.briggsae* homolog is unrecognizable; or may have been assembled de novo via chromosomal rearrangements in the *C.elegans*

genome (Long, 2001). Duplications, chromosomal rearrangements and transposable elements are known to play a role in the birth of novel genes (Betran and Long, 2002; Ganko et al., 2003; Long, 2001). One of the questions that I wanted to address was exactly how important the contribution of transposable elements to novel gene creation was. Specifically, I was able to look at the gene structure level and found evidence of elements creating introns of genes. I used a comparative genomic approach to define this intron creation, regarding such evolutionary questions as whether the *C.briggsae* ortholog also contained an element or intron at this location. Analysis of this subset of elements in introns will enable further insight into the creation of novel genes in general between these two closely related species.

Two families of transposable elements have been identified in *C.briggsae*, termed Tcb1 and Tcb2, both of which are similar to Tc1 in *C.elegans*. Tcb1 (originally called Barney) and Tcb2 were identified by hybridization to a Tc1 probe. (Harris et. al., 1990). The ORFs of Tcb1 and Tcb2 share identity with a structurally similar family of elements named HB found in *Drosophila* (Harris et.al., 1988).

The genomic copy number of Tcb1 and Tcb2 families is 15 and 22 respectively. (in *C.briggsae* strain G16) (Harris et.al., 1990). Two members of the Tcb1 family, Tcb1#5 and Tcb1#10, were sequenced and found to contain an independent single large deletion. Tcb1#5 has a 627-bp internal deletion and Tcb1#10 has

lost 316 bp of one end. A 1616-bp composite Tcb1 element was constructed from Tcb1#5 and Tcb1#10. The composite Tcb1 element has 80-bp terminal inverted repeats with three nucleotide mismatches and two open reading frames (ORFs) on opposite strands (Harris et al., 1990). The composite Tcb1 and the 1610-bp Tc1 share 58% overall nucleotide sequence identity, and the greatest similarity occurs in their ORF1 and inverted repeat termini. Tcb2 is 1606 base pairs in length and contains 80 bp TIRs and a single ORF. For the purposes of this project, I utilized the established sequence for the open reading frame for Tcb1 (X07827) and the complete coding sequence for Tcb2 (M64308).

Regulation of transposition

All *C.elegans* strains contain numerous transposons prone to be mobilized. However, in most strains such as the reference isolate Bristol N2, transposition is only detected in somatic cells but is silenced in the germ line (Emmons and Yesner 1984). In some natural isolates such as the strain Bergerac BO (isolated in Bergerac, France) (Nigon and Dougherty 1949), Tc1 transposons are active in the germ line (Egilmez et al. 1995; Eide and Anderson 1985; Greenwald 1985; Moerman et al. 1986). Bergerac individuals exhibit a mutator phenotype (mut) due to spontaneous mutations caused by de novo Tc1 insertions. EMS-induced mutations of single loci such as mut-2 (Collins et al., 1987) or mut-7 (Ketting et al., 1999; Tabara et al., 1999) are able to activate globally the transposition of multiple Tc families including Tc1, Tc3, Tc4, Tc5, and Tc7.

Transposition silencing in the germ line involves a RNA interference (RNAi)-related mechanism. This connection emerged after the realization that a set of mutants including *rde-2/mut-8*, *rde-3/mut-2*, *mut-7*, -14, 15, and -16 are defective for both RNAi and germ-line silencing of transposition (Chen et al., 2005; Collins et al., 1987; Ketting et al., 1999; Tabara et al., 1999; Tijsterman et al., 2002; Tops et al., 2005; Vastenhouw et al., 2003). I will next detail a bit on what is currently known regarding the connections between transposons, RNAi, and heterochromatin.

RNAi

RNA interference (RNAi) has been found to exist in all organisms studied to date. It was first discovered in plants, where it was termed post-transcriptional gene silencing (PTGS) (Waterhouse et al., 1998). In a simplified model of RNAi, a double-stranded RNA (dsRNA) molecule is cleaved into 21-24 nucleotide-long short-interfering RNAs (siRNAs) by the RNase III-like enzyme DCR-1 of the dicer family. siRNAs are loaded into the RNA-induced silencing complex (RISC) and used for specific cleavage of target RNAs.

Double-stranded RNAs derived from Tc1, Tc3 and Tc5 Terminal Inverted Repeats (TIRs) are indeed detected in the Bristol N2 strain that might arise from the fold-back of transcripts encompassing entire Tc elements. Additionally, siRNAs corresponding to Tc1 and other transposons are also produced in this

strain (Ambros et al., 2003; Sijen and Plasterk, 2003). These siRNAs seem to be functional in the germ line since a Tc1 TIR fused to GFP causes silencing of GFP expression, at least in part, by post-transcriptional silencing of the transgene in the germ line (Sijen and Plasterk, 2003).

Therefore, in this model, RNAi might repress transposition by causing the degradation of transposon-derived mRNA in the germ line, preventing the expression of any Tc transposase. In other tissues, transposon-induced RNAi might be less efficient, thereby enabling somatic excision. However, mutator strains exist that are not RNAi deficient. In *mut-4*, *-5* and *-6* mutant backgrounds, transposition of Tc1 but not of other TC's is specifically derepressed (Mori et al., 1988). These loci have not been identified at the molecular level but they have been proposed to correspond to specific Tc1 copies. For example, truncated Tc1 elements might produce transcripts which lack a sequence targeted by the RNAi system but could still produce a functional transposase. Another explanation could be that these elements might be full-length Tc1 elements inserted in genomic regions that lead to very efficient transcription of these copies in the germ line, hence allowing some transcripts to escape degradation.

In addition to Tc1-specific mutators, a number of genes are required for global silencing of transposition but not for RNAi (Ketting et al., 1999; Vastenhouw et al., 2003). It is not clear if these genes act in a specific branch of an RNAi-

dependent process or if they are involved in an RNAi-independent control of transposition.

A more detailed model of this classical RNAi pathway has been elucidated, including several of the proteins known to be involved in this process. (see Figure 3). It should also be noted that the beginning piece of dsRNA that feeds into this pathway can come from a variety of sources, including; exogenous dsRNA (as is used in RNAi knockdown experiments), endogenous mRNAs, transposons (as described above), RNA viruses, or heterochromatic DNA (Ambros et al., 2003; Bartel, 2004).

RNAi is also involved in the regulation of translation, in which endogenous microRNA precursors (pre-miRNAs) are sequentially processed by the Drosha and Dicer RNase III enzymes, yielding microRNAs (miRNAs). miRNAs bind the 3'-UTR of their target genes and inhibit translation by a currently unknown mechanism (for review, see Carmell and Hannon 2004; Cullen 2004). Various studies have found that many miRNAs are encoded in the genome in a variety of organisms, ranging from viruses to plants to mammals (for review, see He and Hannon 2004; Pfeffer et al., 2004).

While miRNAs are genomically encoded and siRNAs are produced in this process from a variety of sources (see above), both are incorporated into the RISC complex. It is further thought that each incorporate into slightly different

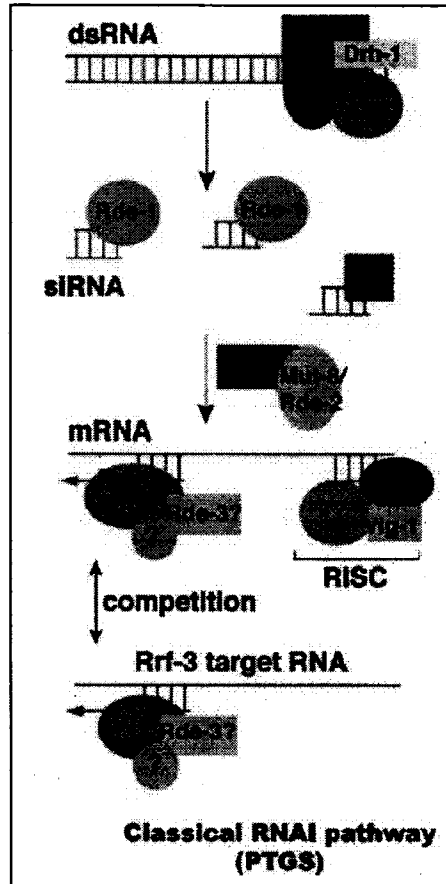


Figure 3 RNAi mechanisms in *C.elegans*

Schematic representation of RNAi mechanisms in *C.elegans*. The classical RNAi pathway is induced by exogenous dsRNA (from synthetic dsRNA (RNAi knockdown experiments), transgenes, inverted repeats, or RNA viruses) that is processed into siRNAs by the Dicer complex containing Dicer, the dsRNA binding protein Rde-4, the PAZ-PIWI protein Rde-1 and the Dicer related helicase Drh-1. Rde-1 is bound to siRNAs bringing them to the next step in the RNAi pathway. Eri-1 antagonizes RNAi by degrading siRNAs. A complex containing Mut-7 and Mut-8/Rde-2 mediates transition between the two steps in the RNAi process. At the downstream step the RISC complex containing a single-stranded siRNA, the PAZ-PIWI protein, Tsn-1 and Vig-1 is shown targeting a mature mRNA. At the same time another complex containing a RdRP (either Rrf-1 or Ego-1) and possibly Rde-3 is engaged in the target-dependent amplification of the dsRNA. A similar complex containing Rrf-3 is shown amplifying its target and creating competition to RdRP complexes involved in RNAi. Reprinted from Grishok et. al. (2005) with permission.

RISC complexes; siRISC and miRISC, respectively (Lee et al., 2004). In siRISC, the target mRNA is silenced by degradation, and in miRISC, the target mRNA is translationally repressed.

Heterochromatin silencing

In addition to the connection between transposons and RNAi, it is also thought that there might be a connection between transposons, RNAi, and heterochromatin silencing. This triple connection is very exciting and has been an active area of my research interest. Heterochromatin, a complex of DNA and associated proteins called histones, was found to possess the ability to silence genes many years ago (as reviewed in Kelly and Fire 1998). Heterochromatin is composed of DNA sequences with little or no coding potential, repeated thousands of times, and silenced by the covalent modification of the DNA itself and of the histones around which the DNA is wound, thus it can be thought of as inactive DNA (as opposed to euchromatin, or active DNA).

Formation of heterochromatin depends on the processing of repeat RNA transcripts into short interfering RNAs (siRNAs), which then direct this formation. For example, siRNAs targeting plant promoters have been shown to cause DNA methylation at these promoters and transcriptional silencing of the corresponding gene (Fire et.al., 1998). More recently, siRNAs have been associated with methylation of histone proteins at centromeric regions in fission yeast--a phenomenon that may lead to silencing of transposons present at the centromeres (Volpe et al., 2002). The hallmark of chromatin associated with silencing is methylation of histone H3 at lysine9 (H3K9), while methylation of H3

at lysine 4 is associated with active genes. Studies in the fission yeast *S. pombe*, *Drosophila*, and plants have connected RNAi related processes with H3K9 methylation. The mechanism of siRNA-mediated DNA and histone methylation is not well understood, but may involve siRNA-mediated binding and guidance of methyltransferases to specific DNA regions. The fact that these siRNAs (which can be derived from transposons) seem to be involved in several of these processes, and that the mechanisms are conserved in a diverse array of organisms makes it a very promising target for discovering how the mechanisms for RNAi, heterochromatin formation and transposition may all be connected.

These results establish a mechanistic connection between RNAi, heterochromatin, and transposons. This connection suggests a possible way in which transposons may be involved in the regulation of host gene activity as well. The regulation of the host genome by transposable elements, which are targets for RNAi mediated chromatin modification and consequent transcriptional silencing of host genes in the region might be subject to this control. This model leads to several testable predictions, and a first step to address these hypotheses requires a comprehensive and global view of the resident sites of each member of each transposable element family.

Fortunately, the genome of *C. elegans* has been completely sequenced and annotated and provides an opportunity to accomplish this task. Furthermore, since a draft sequence of the genome of *C. briggsae* (and soon other nematodes)

is also available, it offers additional opportunities for a comparative genomic approach to these questions. This thesis presents these and other results in an effort to elucidate the answers to this question of potential functional roles of transposons in the genomes they inhabit.

CHAPTER II

METHODS

Bioinformatics – Locations and annotation of elements

Using published sequences for each transposable element family (Tc 1-10, except Tc8) (see Table 3), I performed BLAST searches against both the *C.elegans* and *C.briggsae* genomes located on WormBase (WormBase web site, <http://www.WormBase.org>, releases WS156 and WS157, 2005-2006). Returns with $\geq 90\%$ identities/positives (both values were identical for each hit) were classified as significant hits. All significant hits (regardless of redundancy) were assigned a unique name. All non-redundant hits were identified and annotated in separate files and uploaded onto WormBase to assess research questions. Not all redundant hits (duplicates based on either strand direction or overlapping clone segments) were annotated, so finished annotated files represent all unique hits for elements. Further, all element fragments were identified using published information regarding their characteristics, as displayed in Table 2.

Screenshots were taken of all 40kb regions for each element, and this region was used for data mining concerning questions of number of genes in area,

whether there was an operon in this region, greater than one member of a transposable element (TE) family and assessing number of *C.briggsae* alignments (for *C.elegans* hits). A complete file of all screenshots can be found on the accompanying enclosed CD.

Concerning number of genes in area, this count was based on all protein-coding genes, as denoted by colored (pink/green) boxes in WormBase. Splice variants of a protein-coding gene were not separately included in this count (i.e. 5 splice variants of one gene were counted as one gene). 5 kb screenshots of each region were also captured (and complete file of these screenshots can be found on accompanying enclosed CD) and used to assess characteristics of each fragment (2 ir's, frag) and location of element with respect to genes they aligned with (in introns, exons, whole gene, in transposon annotated gene or other type of gene). For calculations of gene density and classification therein, only protein coding genes not annotated as a transposon or transposon cds were used.

Below is a description of all abbreviations used in data tables.

Region – 40kb surrounding location of element in WormBase. All care was taken to place this element in the center of this region whenever possible.

Type of element – Each transposable element family, as well as the individual type of element was annotated as follows, where X is the number of the transposable element family (1,2,3,4,4V,5,6,7,9,10A,10B) respectively:

Table 2 Abbreviations used

Abbreviation	Indicates a hit that aligns to...
TCXFull	full element
TCX_LIR	left terminal inverted repeat region
TCX_RIR	right terminal inverted repeat region
TCX_T	transposase region
TCX_IS	an internal sequence region (may also see designation IS1, IS2, where 1 and 2 are different internal regions respectively)
TCXPL	left portion
TCXPR	right portion

For display purposes in global linkage group maps, approximate starting position of element (full or fragment) was used, and each corresponding line is an approximate of this position.

Table 3 Tc element sequence source list

Tc	Sequence
1	X01005
2	X59156
3	from John Collins (pers.comm.)
4	GI 156456
4v	L00665
5	Z35400
6	X55356
7	from Reszhohazy et. al. (1997)
9	from Sylvia Fischer (pers. comm.)
10a	from Sylvia Fischer (pers. comm.)
10b	from Sylvia Fischer (pers. comm.)
B1	X07827 – DNA for ORF - GenBank
B2	M64308 – complete cds - GenBank

Sequence Analysis

Unless otherwise noted, all analysis was performed using available tools on WormBase (www.WormBase.org), Biology Workbench 3.2 (<http://workbench.sdsc.edu/>) and Ensembl (www.ensembl.org) using all default parameters.

For the intron study portion of this research, the following procedure was followed: For either a *C.elegans* or *C.briggsae* gene for which a transposable element aligned (described as E Intron Study and B Intron Study, respectively), an alignment of the element and the gene was first performed using ALIGN on Biology Workbench. Corrections for directionality were made, such that alignments were not biased for direction. Additional alignments (using CLUSTALW, ALIGN, and CLUSTALWPROF on Biology WorkBench) were also performed with both genomic and cDNA sequence to assess the complete location (within intron, exon, combination) of the transposable element within the gene.

Orthologs of genes were chosen first by the ortholog given in WormBase, and in instances where no ortholog was listed; the Best BLASTP match gene was used in these studies (and is noted where applicable). Additionally, the Synteny Viewer was also utilized on WormBase; however the status of this tool and subsequent annotation is questionable, and thus was never used as a sole determinant of results.

Next, for determining whether the ortholog (*C. briggsae* or *C. elegans* respectively) contained the transposable element located in the element intron gene, I used the BLS2SEQ program in WorkBench. Specifically, I took the genomic sequence of the ortholog and the respective transposable element and BLASTed them together to look for any similarity.

Finally, a series of alignments were conducted to address the question of whether the corresponding ortholog contained an intron in the same location as the respective element intron gene.

First, I determined the alignment of the transcript of the element intron gene with its corresponding protein sequence utilizing either readily available alignments on Ensembl (for all *elegans* intron genes) or by producing an alignment using the WISE2 tool on EMBL (<http://www.ebi.ac.uk/Wise2/advanced.html>) (for *briggsae* intron genes). For the *briggsae* genes, I used the available genomic sequence information and protein sequence information as available in WormBase.

Secondly, I produced a texshade alignment of the element intron gene and the corresponding ortholog using the following method. I aligned the element genomic and cDNA (or predicted cDNA for all *briggsae* element genes) sequences for the respective element intron gene using ALIGN on Biology WorkBench. Next I aligned the ortholog genomic and cDNA (or predicted for all

briggsae orthologs) using ALIGN on Biology WorkBench. Subsequent to these two alignments, I performed a CLUSTALWPROF, which produced a multiple alignment of these two pairs of alignments, such that all four sequences were aligned. Finally, I produced a texshade display for each of these multiple alignments, which can be found on the supplemental CD.

Additionally, I performed all of the above alignments with the element intron partially removed, so as to achieve a substantially more effective and accurate alignment (I had previously determined for all of the orthologs that no transposable element was existing in these genes by the BLS2SEQ procedure described above).

All data utilized for this study is archived on my Biology Workbench in sessions E Intron Study and B Intron Study, for analysis of *C.elegans* and *C.briggsae* gene groupings, respectively. For large genes for which CLUSTALW was not possible, I choose segments of relevant genes for the analysis (and these are noted within results, where applicable).

CHAPTER III

RESULTS

Are resident transposable elements functional components of the genomes they inhabit? For example, do the resident transposons serve as targets for cis regulation? If they are in fact regulatory elements, you might expect to see 1) conservation of these elements in closely related species and 2) clusters of genes commonly regulated (as evidenced by similar expression patterns) located near such transposable elements. In order to address this larger research question and look for potential evidence of their functional role, I needed to establish the locations of all the transposable elements in the *C.elegans* genome. In addition, not only did I need to locate where all the full TE's reside, but it also would be necessary to locate if and where there resided fragments of these same elements. Other studies have established where most of the full elements reside (Fischer et al., 2003), but no one to date has published data on where fragments are located. In order to address related questions about the role, function, and evolutionary consequence of transposition in general, the locations of these fragments in addition to full elements was a critical need. In addition to this need to answer my own research questions, it also became apparent that information on the locations of all these elements could be gathered and

organized in such a way as to provide permanent annotation to both the *C.elegans* and *C.briggsae* genome databases, and thus be useful and of value to the entire scientific community.

As described in methods, I performed BLAST searches on both the *C.elegans* and *C.briggsae* genomes using published sequences of transposable elements (see Table 3). For purposes of this project, I wanted to be most conservative in what I called a significant hit, and thus only included those returns of $\geq 90\%$ identity (past published searches have used 80% or greater to signify a hit). While performing these searches, I analyzed the different types of identity returns with regard to the 80-90% identity differences, and the 90% identity seemed to be a significant breaking point where most of the returns were of sufficient length to constitute a significant hit. I also verified that all the full elements I located by this method were the same as previously published, thus assuring any comparisons to previous publications would be relevant. A summary table of all the significant hits I found in this manner can be seen in Table 4. Based on results of this project, I intend in the future to generate another list of those elements with 80-90% identity and determine if any different patterns result, although on an anecdotal note while I was visualizing them, no apparent differences were striking (i.e. there were not a larger number of particular te families' elements – the distribution of full/partial was similar as well).

From these significant hits, I located genomic coordinates for all, and created annotation files for each. These annotation files can be found on the supplemental CD, and are ordered by TE family (Tc1, 2, etc.). As detailed in the methods section, annotation files contain only unique hits, and can be directly uploaded into WormBase in order to visualize each element alongside all the characteristics already available in WormBase. In this way, I was able to ask questions regarding the position of elements, both full and partial, located throughout the *C.elegans* and *C.briggsae* genomes.

Global overview of locations of transposable elements in *C.elegans*

I located and annotated 276 elements (both full and partial), 84 of which were full-length elements, and 192 were partial fragments, heretofore unpublished, or annotated (See Table 4). Of the 84 full-length elements, 69 were elements of TE families evidenced to display transposition activity (Tc's 1, 2, 3, 4, 4V, 5 and 7, as detailed in the introduction). Of the 192 partial fragments, 119 of these were of elements of active TE families.

Location by TE family

With respect to the element families, I located only 1 full copy of TC's 10A and 10B from the inactive element families and only 4 full copies of Tc2 from the active families. In contrast, I located 27 full copies of Tc1, which is the number of

these full elements previously described. Regarding partial fragments, the lowest and highest counts were for Tc2 (0), and Tc4V (57), respectively.

Table 4 Locations of all Transposable Elements
GLOBAL ANALYSIS OF ELEMENT LOCATIONS

Element	Linkage Group						TOTAL
	I	II	III	IV	V	X	
1F	3	7	2	3	10	2	27
1P	0	3	0	0	3	0	6
2F	0	1	0	0	3	0	4
2P	0	0	0	0	0	0	0
3F	5	7	2	2	2	0	18
3P	0	0	0	2	2	7	11
4F	0	1	1	0	0	1(s)	3
4P	9	0	3	2	6	9	29
4VF	3	0	0	0	1	0	4
4VP	9	2	5	6	11	24	57
5F	0	1	0	2	0	0	3
5P	1	4	0	6	2	2	15
6F	0	4	2	2	3	2	13
6P	4	3	4	2	9	1	23
7F	0	1	0	1	2	6	10
7P	0	0	0	0	0	1	1
9F	0	0	0	0	1	0	1
9P	10	1	0	10	11	12	44
10AF	0	1	0	0	0	0	1
10AP	1	1	1	1	0	3	7
10BF	0	0	0	0	0	1	1
10BP	0	0	0	2	0	0	2
Full	11	23	7	10	22	11	84
Partial	34	14	13	31	44	56	192
Full Active	11	18	5	8	18	9	69
Full Inactive	0	5	2	2	4	2	15
Partial Active	19	9	8	16	24	43	119
Partial Inactive	15	5	5	15	20	13	73
Grand TOTALS	135	111	60	123	198	67	276

Concerning each element family, the general trend is a lack of one. For example, Tc1 elements, both full and partial, appear scattered throughout each linkage group, and appear on each linkage group. There are some cases where a Tc family does not have copies on every linkage group (Tc6 full elements are found on all except Linkage Group I), but this doesn't appear to be significant, as the total number of elements for these examples is low to begin with (Tc6 has only 13 full elements).

Distribution of Fragment elements

Another novel question regarding distribution of elements came out of this analysis, that being what patterns existed concerning the fragments of elements found. There were no fragments found for Tc2, and thus it was not analyzed for this portion. Additionally, since Tc's 9, 10a, and 10b were predicted solely based on genomic searches (and there is not data regarding their respective portions of elements – ir's, transposase, specific internal sequences), they are left out of this portion of analysis as well.

One of the questions I wanted to address was how the fragments were split up with regards to each element family's specific architecture – do we locate most fragments resembling the inverted repeat regions, the transposase gene, a combination of the two? Additionally, how are these fragments distributed throughout the genome – do we locate particular subsets of types of fragments localized to a linkage group? Table 5 provides a summary of data to help elucidate answers to these and other questions.

Table 5 Distribution of element fragments

	1IR	T	COMBO	IRS GAP		
LG						
I	7	2	3	10		
II	1	0	2	2		
III	1	0	3	4		
IV	6	0	5	4		
V	6	0	4	8		
X	6	0	1	10		
TE*					FRAG	FRAGMENT TOTALS FOUND
1	0	0	2	0	4	6
3	1	0	1	0	9	11
4	11	1	3	10	25	29
4V	3	1	5	22	26	57
5	7	0	1	4	3	15
6	5	0	5	2	11	23
7	0	0	1	0	0	1
TOTAL	27	2	18	38	78	
<p>Analysis of fragment distribution, by linkage group (top of chart) and by te family (bottom of chart). 1 IR – a single fragment that matches one inverted repeat end. T- a single fragment that matched only the transposase section. COMBO – a single fragment that matches a combination of IR, T, and/or IS. IRS GAP – two fragments, each counted separately for consistency, each of which corresponds to an IR respectively, with a gap between them in genomic sequence. FRAG – two or more fragments that correspond to portions of a transposon (each counted separately for consistency) that have fragments that overlap, but do not compose a full element. *TC 2, 9, 10A, 10B not included in this analysis, as there were not fragments found for Tc2, and 9, 10A and 10B do not have data regarding IR, T sections.</p>						

Several patterns of distribution are apparent from this data. The trend is clearly to locate fragmented elements of all kinds except just the transposase gene. Additionally, there do appear to be a significant number of ir's (38 – which represents 19 pairs of IRS) remaining in the genome that no longer contain the transposase gene (IRS GAP in table and see Figure 4). One question to address would be what is in fact now located in the sequence between the ir's, and is one I will be pursuing (outside of the scope of this dissertation).

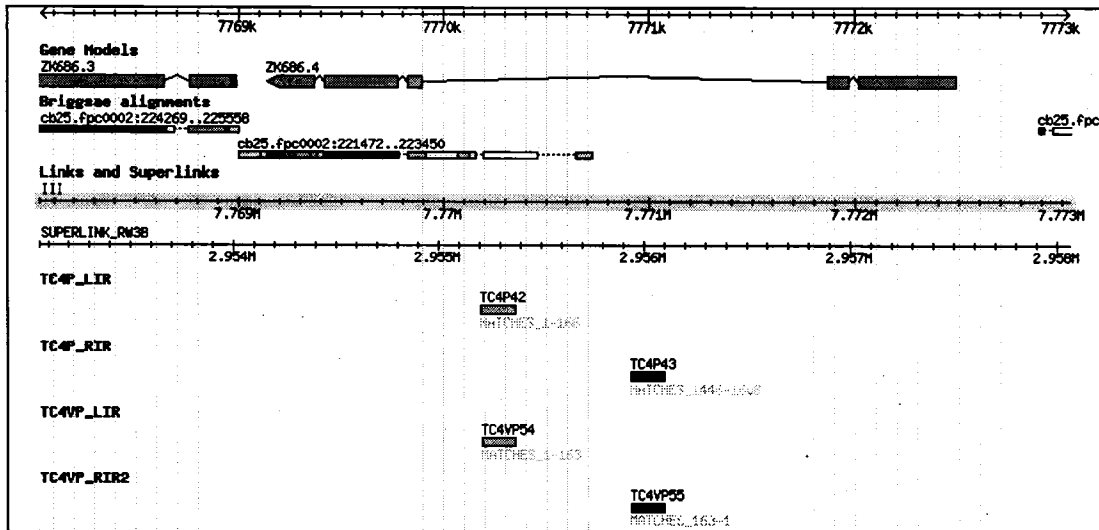


Figure 4 Example of IRs with GAP

I also found a number of fragments (78 – denoted FRAG and can be seen in Figure 5), those being parts of elements that appear to overlap each other, but do not form a full element. Keep in mind that this number reflects each fragment – thus the total number of actual sites of fragments is considerably lower (78 represents 27 sites). Presumably these are past active elements that have undergone mutation such that portions of the sequence corresponding to the element have been deleted. As with the fragments that appear like a pair or ir's with sequence of unknown origin between, this subset awaits further investigation.

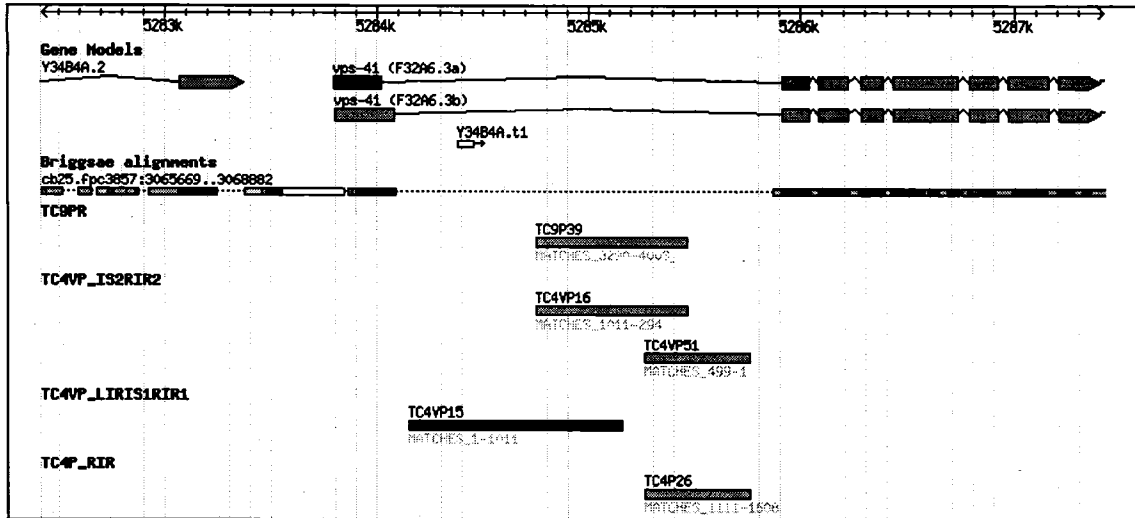


Figure 5 Example of FRAG

As described in the introduction, Tc 4 and Tc 4v are directly related to one another in the sense that portions of both are the same and 4v contains an extra inserted region. Thus, I would predict that the locations of fragments of each that I found would correspond to one another, and they in fact do. All of the regions with Tc 4 fragments also contain Tc 4v fragments, and the patterning of fragments matches what is already known about the architecture of each (i.e. Tc 4 IR regions match to Tc 4v). Additionally, Tc 9 has direct sequence alignment with portions of both Tc 4 and Tc 4v, and you do see these portions of aligned sequences resulting in overlapping fragments in this study. A screenshot representative example of this clustering can be seen in Figure 6.

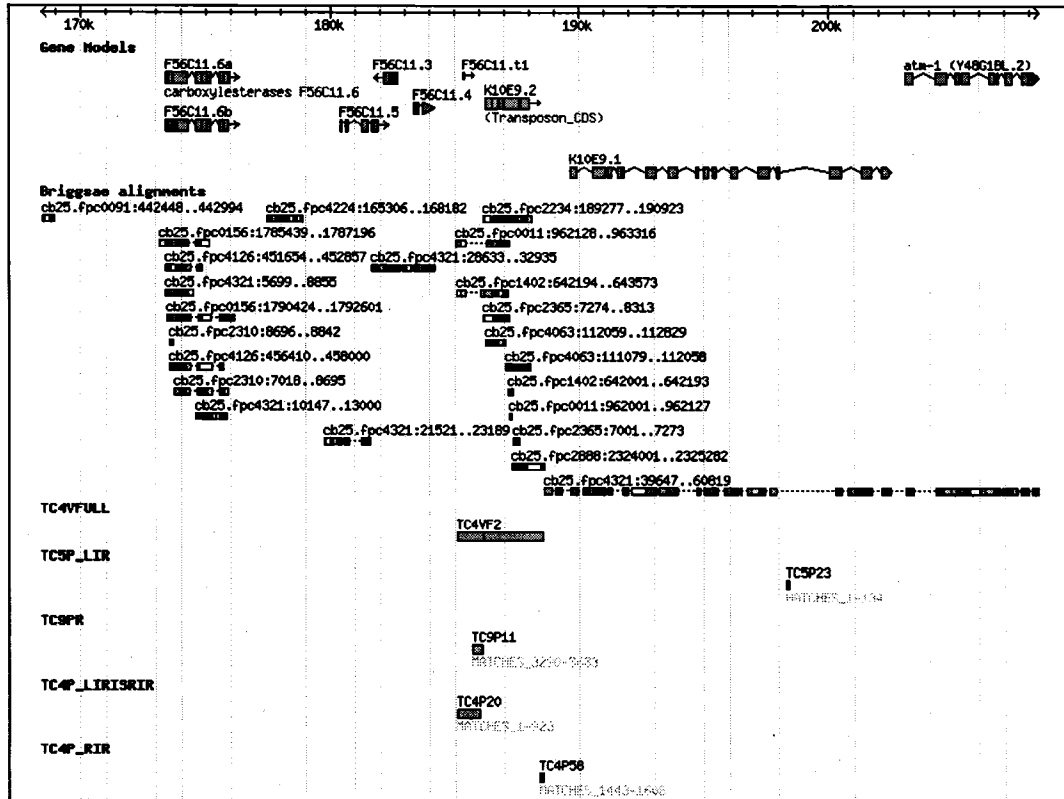


Figure 6 Screenshot of Tc 4, 4v, 5, 9

Location within each Linkage Group

Another question I wanted to address was the location of the elements by chromosome. It had been reported previously that full elements were primarily located on the ends of all the linkage groups, with very little distributed throughout the central portions of linkage groups (Fischer et al., 2003). Additionally, a correlation has been found between the density of DNA transposons and the regions of higher chromosomal recombination rate (Duret et al. 2000; Rizzon et al. 2003). As mentioned previously, there have been no published data on locations of fragments of elements. Some of my results are in

contrast with several areas previously reported, as can be seen in Table 4 and Table 5, as described below.

Regarding trends in overall global position of elements (which linkage group they reside on), I found that LG III had the lowest number of full elements (7) and LG II to have the highest number (23). This finding is in keeping with previous reports of LG III containing few full elements (Fischer et.al, 2003). Additionally, LG II and III appear to have a comparatively low number of partial elements (13 and 14) with respect to the rest of the linkage groups (34, 31, 44, and 50).

In addition to understanding where these elements reside on a global scale, I also wanted to determine at a finer scale what patterns might exist where elements are located within linkage groups (ends/center) and even at a more detailed scale, where they reside with respect to other elements of the genome (intergenic, in introns/exons, regions of gene rich/poor, etc.) On the enclosed supplemental CD is a tabulated version of these results, and I will highlight a few key locations from this analysis below.

To better visualize where each element resides with respect to position within a linkage group, I created several compilations of this data. Appendix A is one of these compilations, where I utilized the genetic maps for each linkage group, and approximated the location of every element. In addition, I color coded each element so that you can see not only location by each element family, but also in

several other groupings (all active elements, all partial elements, overall total elements, etc.).

From this depiction and tabulated data, several trends were obvious. The overall pattern again was actually a lack of one. For the most part, all the elements appear scattered throughout each linkage group (no association with locating more at the ends as was previously described) as well as across all linkage groups (there doesn't appear to be any one linkage group that is in stark contrast to the other linkage groups with respect to total locations). With that said, there do appear to be several localized clusters of fragments on Linkage Groups V and X, both somewhat near the ends of each linkage group.

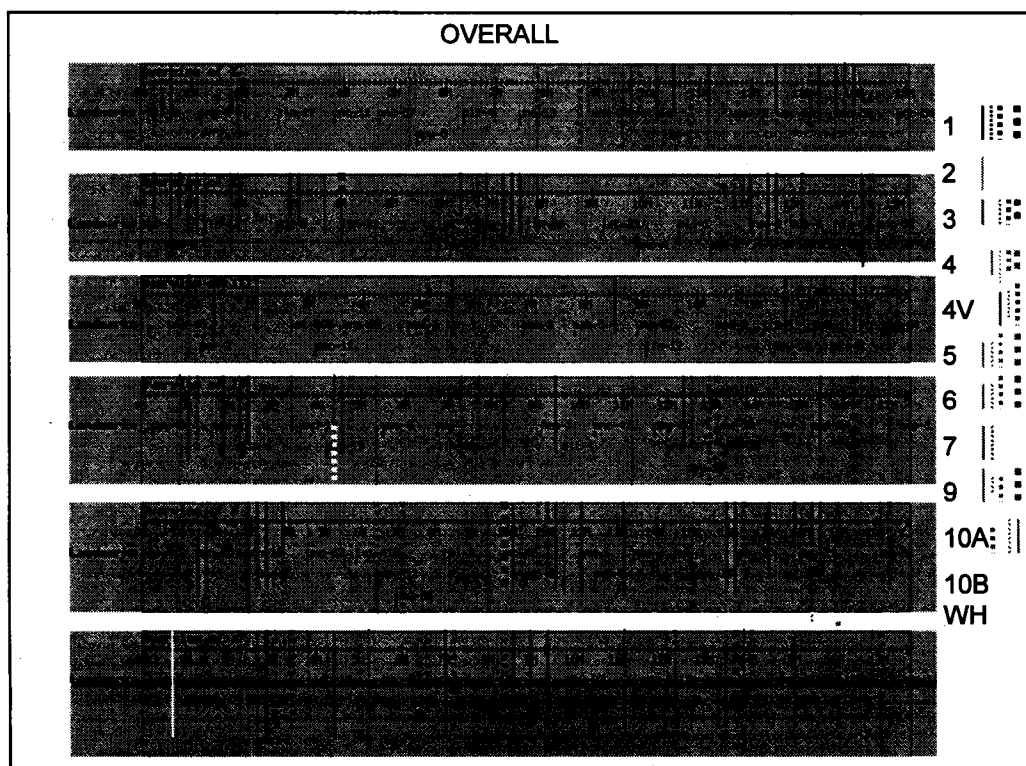


Figure 7 Overall position of all TE's

Also related to global position of elements, I was intrigued by the previously published idea that transposons were found primarily in gene poor regions, as previous studies reported the majority of the full elements to be located on the chromosomal arms. In one of these previous analyses (Fischer et. al., 2003), all of the full transposable elements were analyzed and located by BLAST searching using published sequence information (see Figure 8 below).

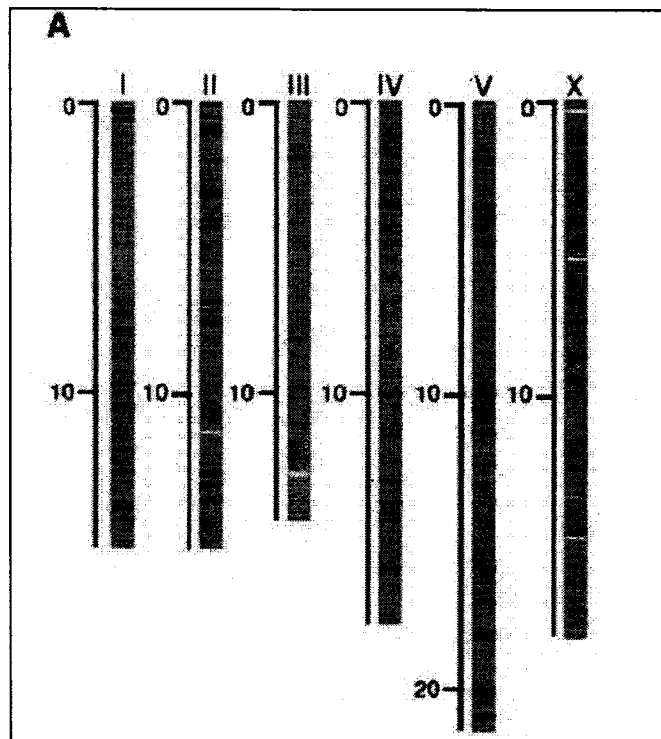


Figure 8 Locations of Active Transposable Elements (Fischer et. al. 2003)
 (A) Genomic distribution of all active transposable elements in the *C. elegans* genome. The positions of the transposons on the physical map are based on the positions of the clones annotated in WormBase. The sizes of the chromosomes are in megabases. The transposons are color coded as follows: Tc1, blue; Tc2, green; Tc3, red; Tc4, white; Tc4v, black; Tc5, light blue; and Tc7, yellow. Reprinted from Fischer et. al. (2003) with permission.

It was stated in this publication that relatively more transposable elements are found on the autosomal arms than in the middle third of the chromosome. This statement was based on the global gene density that has been calculated (see

Table 7), and thus can only correlate gene density and locations of transposable elements in a very broad based sense (i.e. more elements are located on the arms, therefore located in "gene poor" areas). I was interested in further analyzing this potential correlation, and thus determined the gene densities on a much more specific scale for each region for where a transposable element was located.

For this fine scale analysis of gene densities of transposon regions, I computed the totals of protein coding genes located in each 40 kb region and then converted to represent gene density in gene/Mb. Table 6 displays the result of this calculation for my results (raw data used for this analysis can be found on the enclosed supplemental CD).

Table 6 Calculated Gene Density
Gene Density of Transposon Regions

LG	TOTAL 40Kb REGIONS	TOTAL GENES	GENES/Mb
I	23	199	216
II	30	329	274
III	12	104	217
IV	25	211	211
V	36	414	288
X	34	263	193

Genes/Mb based on (Total Genes/Total Regions)/40Kb
(1000Kb/1Mb)

Comparing this calculated gene density of transposon regions with previously published data regarding gene density on a global scale for the worm genome

(see Table 7 below); you can see that there is not a predilection to locate transposons in gene poor areas as was previously described. In fact, the gene

Table 7 Table 1 repeated¹

Chromosome	Size (Mb)	Protein-coding genes	Density (genes/Mb)
I	15.08	3260	216
Left	4.00	685	171
Center	6.26	1573	251
Right	4.82	1002	202
II	15.28	3874	253
Left	5.90	1648	279
Center	5.44	1435	263
Right	3.94	791	201
III	13.76	3103	225
Left	4.80	972	202
Center	4.29	1199	279
Right	4.68	932	199
IV	17.49	3606	206
Left	6.74	1339	198
Center	5.08	1321	260
Right	5.67	946	167
V	20.92	5256	251
Left	6.51	1615	248
Center	6.99	1880	269
Right	7.42	1761	237
X	17.72	3186	180
¹ Adapted from Spieth et.al, (2006) with permission.			

density for transposon locations on linkage groups V and X is greater than the global genome density of these regions.

Genetic position (intergenic, in introns, exons, etc.) of elements

Another subset of questions that are more detailed was addressed in my analysis of each element. The first of these was where each element was located with respect to genetic position – intergenic, intragenic (in introns or exons?), etc. To reiterate, my guiding question/hypothesis in this research is the possibility that transposable elements have a functional role in the *C.elegans* genome, and thus their relative positions can help to elucidate answers to this possibility.

Table 8 Tabulation of genetic positions of TE's

TOTAL REGIONS	TOTAL GENE REGIONS	TOTAL GENES*	TOTAL FULL TE IN GENES	TOTAL PARTIAL TE IN GENES
160	65	68	34	90
NOTE: GENE CLASSIFIED AS A PROTEIN CODING, NON-TRANSPOSON GENE				
*OF TOTAL GENES, TE ENCOMPASSES...				
WHOLE GENE	I/E COMBO	EXON	INTRON	
3	3	1	61	

As can be seen in Table 8 above, transposable elements are almost equally likely to be found in gene regions (41%) or intergenically (59%), a surprising result based upon the currently held idea that elements would be primarily found in gene poor regions. Not surprisingly, of the elements found within genes, the vast majority of them are found in introns (61 out of 68 total genes). Interestingly, there are very few (3 of 68) examples in which elements found within genes encompass a combination of introns and exons together (denoted I/E combo. in table below). One of the elements resides completely within an exon. This

element could be a transposase not yet annotated as such and awaits further investigation. The remaining three elements comprise whole genes, and are presumably transposons not yet annotated as such in WormBase. These results led me to investigate further into the subgroup of elements that encompass introns (see Intron Studies beginning on page 51).

Operons and transposable elements

As mentioned in the introduction, *C.elegans* (and *C.briggsae*) are unique (among eukaryotes) in that they have genes organized into operons. As I was analyzing the locations of all the transposable element hits, it appeared like there were a lot of operons in the regions I was looking at, which led me to go back and systematically categorize each 40kb region where an element resided as containing an operon in that same region. Of 160 regions investigated in this project, 55 were regions where operons we also found (Note that the operon did not in fact have to cover the same area of sequence as the element in this analysis). There are about 1000 operons in total in the *C.elegans* genome, and 96% of these are conserved in *C.briggsae* (Stein et. al., 2003) It has been previously established that 15% of *C.elegans* genes are part of operons and these operons are concentrated in the central clusters of *C.elegans* chromosomes (Blumenthal et. al. 2002). Thus, there does not appear to be any correlation to presence of a transposable element and presence of an operon.

Comparative Genomics – *C.elegans* and *C.briggsae*

C.briggsae, as mentioned previously, is a closely related species to *C.elegans*, and its genome is currently being assembled and annotated. It is possible to compare the sequences of these two nematodes, and thus, one of the questions I wanted to address was what were the similarities/differences that existed between these two genomes with regards to transposable elements. This comparison is a natural result from my research hypothesis that transposable elements serve a functional role. An expectation of such a hypothesis would be that similar species would exhibit conserved regulatory elements, and thus, you might expect to find a similar patterning of transposable elements in this closely related species.

A first run through to address this question in a general sense was achieved by totaling the number of *C.briggsae* alignments displayed in WormBase for each element I found (see data tables on the enclosed supplemental CD). In order to identify potential regions to address questions regarding synteny of that region, I further categorized each region as exhibiting synteny as denoted as having >10 *C.briggsae* alignments in that region. This part of my analysis was for identifications of potential regions of synteny only, and not meant as a quantitative study. This subset of regions exhibiting potential *C.briggsae* synteny remains to be further investigated. Some questions to be addressed are what resides in these gene regions in *C.briggsae*, is there evidence of an element, either currently residing, or evidence of a past insertion/deletion? One piece of

evidence to look for would be presence of the footprints left behind when a transposable element is cut from a position. Unfortunately the *C.briggsae* genome (and current Synteny Viewer of WormBase) is not in a sufficiently completely assembled and annotated form to address these questions in an efficient manner, but they are nonetheless important questions for understanding the evolution of these two species as it relates to transposition.

Transposable element presence

I was able to address the general question of whether *C.briggsae* had transposable elements and other characteristics regarding their locations. Two elements had been previously discovered in *C.briggsae* (Harris et. al. 1988, 1990; Prasad et. al. 1991), and both are related to the Tc1 element in *C.elegans*. They are termed Tcb1 and Tcb2. Along with BLAST searches of the *C.briggsae* genome with these published elements, I also performed searches with all the Tc 1-10 (except 8) elements (listed at Tc1b-10b for purposes of this analysis) as described here for *C.elegans*.

Table 9 tallies the results from these BLAST searches. While I did locate significant fragment matches for all families, I only located/annotated full matches for this project. Tc families 3, 4, 5, 7, 9, and 10b did not return any significant full hits corresponding to *C.briggsae* elements of these families. I do see the same pattern as in *C.elegans* of a subset of full elements (Tcb1 and Tcb2) that match up to introns of genes. This subset of elements (with the exception of TcB1F19,

TcB2F13, and TcB2F14 due to size constraints on corresponding genes) and corresponding genes were used in the B Intron study, described below (and for which all supplemental data is located on the CD in the B intron study folder).

Thus, 4 Tcb1 elements and 7 Tcb2 elements were further investigated.

Additionally, I did locate full elements with similarity to *C.elegans* Tc's 2, 4V, 6, and 10A. Surprisingly, I located more Tc2 related elements in *C.briggsae* than I did in *C.elegans* (16 vs. 2).

Table 9 *C.briggsae* full elements

Full Element	Total	No Gene	Intron	Part Gene	N/A
TcB1	19	10	5	1	3
TcB2	41	28	9	3	1
Tc2B	16				
Tc4vB	3				
Tc6B	2				
Tc10AB	3				

Intron Studies

One area of comparison between these two genomes that I was able to pursue (given the current annotation status of the *C.briggsae* genome) involved looking more closely at the subset of *C.elegans* transposable elements that appeared to encompass entire introns of genes. Fifteen of these element introns were analyzed which comprised all the full elements which appeared to encompass

most or all of an intron in a *C.elegans* gene. Additionally, I also did these described analyses starting with the full *C.briggsae* elements (4 of Tcb1 elements and 7 of Tcb2 elements as described above). These two analyses are herein described as E Intron Study and B Intron Study, respectively. For both studies, no element was found to be contained within the orthologous gene (i.e. for an *elegans* Tc1 intron element gene, the *briggsae* orthologous gene did not contain a Tc1 element).

Additionally, from BLAST searches and annotations, none of the orthologs of these gene pairs contained elements from any other Tc families (1-7, 9, 10). Briefly, since I had annotated all the full and partial fragments for *C.elegans*, any of the *elegans* orthologs were quickly examined (by visualizing each gene in WormBase) to ascertain that they did not contain any full or partial element. For *C.briggsae*, I conducted BLAST searches using all the published sequences for all the elements (as described above), so I cross checked this list against any *C.briggsae* ortholog to determine that no full fragment existed (Excel charts of all the elements found in this manner are located on the supplemental CD). I was not able at this time to ascertain whether any of the *C.briggsae* orthologs contained fragments of Tc's 1-10, as I did not keep a record of these fragment hits.

C.elegans Intron Study (E Intron Study)

For the full *C.elegans* elements that encompassed most or all of an intron of a gene (15 of 34 full elements found in introns – the remaining did not encompass close to the full intron of their respective gene), I analyzed each to determine the exact position of the element within the intron as well as looking at relative similarity between *C.elegans* and *C.briggsae*. Briefly, I took the sequence of each respective element, the corresponding gene, and performed a global sequence alignment (using ALIGN on Biology Workbench). All of these alignments and subsequent texshade images I produced can be found on the supplemental CD.

To gather information regarding the relative similarity in alignments between the *C.elegans* gene containing an element intron and the corresponding *C.briggsae* ortholog or best BLASTP hit (where an ortholog was not listed on WormBase), I performed several types of alignments using the available tools on WorkBench, Ensembl and EMBL (<http://workbench.sdsc.edu/>, <http://www.ensembl.org/index.html>, <http://www.ebi.ac.uk/Tools/>).

Specifically, I removed most of the intron (leaving between 50-110 bp on either end of intron) corresponding to where the element was located in *C.elegans*, and did an alignment of this sequence with both the genomic sequence and cDNA sequence of the *C.elegans* gene where possible. For some of the larger genes it was not possible to align all three sequences in this manner for the entire gene, in which case I performed this alignment on a smaller segment of each gene pair.

I then aligned this pair with the pair of aligned *C.briggsae* ortholog sequences (genomic and predicted cDNA sequence formed the pair). Additionally, I was also able to use Ensembl (www.ensembl.org) to gather the transcript and protein information for the *C.elegans* gene in one alignment, where the codons for each amino acid were aligned over one another. For the B intron study, I was also able to perform this same function (aligning transcript and corresponding protein sequence together) utilizing the Wise2 tool on EMBL (<http://www.ebi.ac.uk/Wise2/>).

By combining both the Ensembl alignments with the above described alignments in WorkBench, I was able to determine where the intron splice sites for each *C.elegans* and *C.briggsae* ortholog were, and answer the question of whether these two genes did in fact have the same exon/intron junction. In addition, to verify this answer, I also had the protein alignments for the *C.elegans/C.briggsae* pair, as well as an alignment of *C.elegans/C.briggsae* genomic/cDNA sequences along with the *C.elegans* DNA with the portion of sequence corresponding to the transposable element removed. An example of the alignment studies performed to address the question of whether there was a conserved intron can be seen in Figure 9 below. Additional images of each alignment for each *C.elegans* "element intron" gene can be found in Appendix B – E Intron Study.

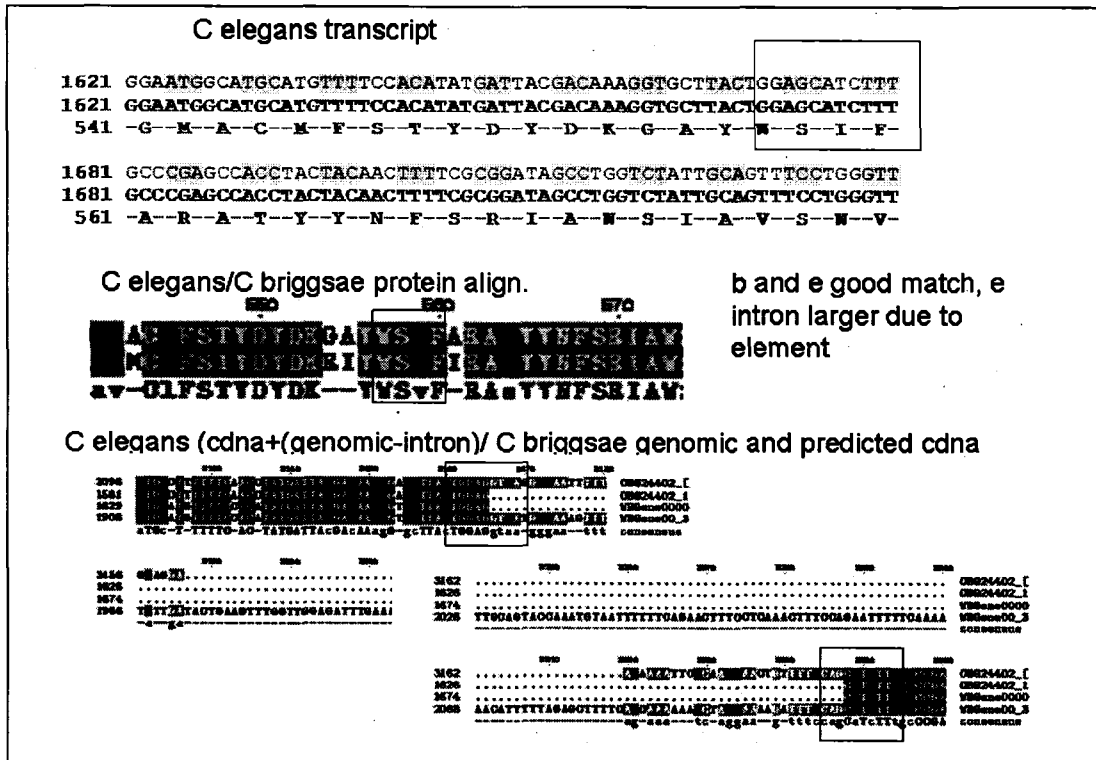


Figure 9 Example of a conserved intron (C31A11.7/CBG24402)

Of the 15 *C. elegans* element introns investigated, four main subgroups of elements were apparent. One group have clearly conserved element introns with respect to their *C. briggsae* ortholog. For the second group, I was unable to answer the question of conserved introns due to either lack of a briggsae ortholog or lack of exon/intron annotation. A third group was unclear with regards to presence of an intron, due to the present annotation of the *C. briggsae* genome (see details below). Intriguingly, the fourth and final group appear to be examples where a transposable element actually created an intron, splitting apart a pre-existing exon into two separate exons.

As mentioned above, five *C.elegans* genes have clearly conserved element introns with respect to their *C.briggsae* ortholog (C31A11.7 (shown above in Figure 9), *sra-28*, *imb-2*, T02G5.11, and F30F8.10, also see Table 11). To clarify, this is where the element intron is located in the *C.elegans* gene and there exists an intron at this location in the *C.briggsae* ortholog. Additionally, the intron in *C.elegans* is much longer than the corresponding intron in *C.briggsae*, due to the presence of the transposable element. These five *C.elegans* genes also have similar overall gene structure to their *C.briggsae* ortholog, with regards to their locations and number of exons and introns. This can be easily visualized by the scaled depictions of each pair I created (all can be found in Appendix B) and verified by the alignments I conducted. It was very clear from both the scaled depictions and the actual alignments that I conducted that these elements had inserted into an existing intron in *C.elegans*. This also provides evidence that these insertions occurred following the evolutionary divergence between *C.elegans* and *C.briggsae*, since no vestige of an element is found in any of the five corresponding *C.briggsae* orthologs.

Table 10 *C.elegans* Intron Study Part I

<i>C.elegans</i> Full Element matches intron				
E Gene	Element*	CBG	Element matches gene	Position between exons
C31A11.7	1F2	24402	2098-3709	7&8 (1734-1952)(4235-4329)
clec-41	7F6*	9432	2158-3080	after exon 3 (final coding exon)
F02D10.6	7F10*	23672	104-1025	1&2(1-54)(1100-1276)
hum-7	3F9	3901	6475-8811	10&11 (6197-6426) (8913-9011)
IMB-2	10AF1	11089	3253-6797	6&7 (2682-2932) (7526-7633)
mdt-29	6F25	18261	1945-3567	4&5 (1648-1878)(3910-4257)
sra-28	1F18	4324	945-2555	2&3 (459-877)&(2601-2725)
srh-291	1F11	4863	314-1924	1&2 (1-310) (2413-2837)
srw-83	3F12	16962	977-3312	3&4 (832-877) (3335-3444)
T02G5.11	3F4*	24744	633-2969	3&4 (492-619)(3004-3157)
T05H4.10	9F1*	18929	2124-6416	8&9 (1790-1939)(6440-6580)
T07D3.3	1F1	7159	349-1959	1&2 (1-273)(2298-2402)
T24E12.10	6F15	4386	2473-4076	12&13 (2249-2428)(4142-4509)
ZK856.5	1F19	9540	866-2475	3&4 (804-849)(2643-2775)
F30F8.10	4VF3	23731	2197-5714	2&3 (1696-1980)(5919-6151)
* indicates element is on opposite strand from gene – reverse direction				

A second group of *C.elegans* element intron genes and their *C.briggsae* orthologs were not able to be assessed with regards to presence of conserved introns. *C.elegans* genes F02D10.6 and T07D3.3 turned out to in fact not have a *briggsae* ortholog (at least at present). For both of these *elegans* genes, no ortholog was listed in WormBase, thus I used the Best BLASTP match as a potentially orthologous gene in this study. Both of these choices had very weak similarity as evidenced by their respective alignments (both protein and nucleic acid), and thus it appears are not orthologous to the *elegans* intron gene. It is of note that there were several other *C.elegans* element intron/*C.briggsae* Best BLASTP pairs (6 out of 8) that did in fact have a good deal of similarity (in fact

three of these; C31A11.7, T02G5.11, and sra-28, clearly had conserved introns), thus utilizing the Best BLASTP match appears to be a valid approach to finding a majority of orthologous genes when one is not specifically listed.

Table 11 *C. elegans* Intron Study Part II

<i>C. elegans</i> intron gene/ <i>C. briggsae</i> ortholog (cbg)*	e/b unspliced gene (bp)	e/b protein (aa)	intron consv.	Additional Notes
C31A11.7/24402*	4853/4167	706/690	Y	
sra-28/04324*	3069/1323	341/348	Y	
imb-2/11089	8132/6579	883/879	Y	had to take out first part of each gene and element of e gene as very large genes
T02G5.11/24744*	3345/1293	184/262	Y	
F30F8.10/23731	6361/3357	245/247	Y	probable syntenic region
clec-41/09432	1638/1733	545/546	N	element made intron – noncoding exon4 aligns to flanking briggsae ortholog sequence
F02D10.6/23672*	2016/3262	199/212	N/D**	spotty protein alignment – b gene not orthologous
hum-7/03901	21741/8574	1887/1890	N	protein alignment dissimilar only in region of intron junction. b has exons and introns in e intron region
mdt-29/18261	5443/1563	441/469	N	b and e similar – e gene much larger with larger introns – could have conserved intron but unable to determine
srh-291/04863*	3126/6579	326/879	N	b and e similar, b has patches of similarity of alignment in e intron, but all appear as exon – possible element created intron in elegans
srw-83/16962*	3958/995	336/281	N	b annotated as having exon extending into same region where e element intron begins
T05H4.10/18929	6658/1757	476/476	N	element created intron in elegans, no intron in briggsae, but good alignment everywhere else, including potential syntenic region
T07D3.3/07159*	3517/2499	284/447	N/D**	b gene not orthologous
T24E12.10/04386*	4509/1785	632/594	N/D**	no introns on b gene model prediction
ZK856.5/09540	5381/3293	755/561	N	ortholog matches part of e gene after element intron
* indicates no ortholog listed in WormBase, thus BEST BLASTP match listed was used ** N/D either no similarity existed between briggsae gene or an ortholog was not available, thus question of intron conservation could not be determined				

The third group of *C. elegans* element intron genes (hum-7, mdt-29, srh-291, srw-83, and zk856.6) all were inconclusive with regards to the question of conserved introns, due to a variety of specifics with a general theme being that of the current state of annotation of the *C. briggsae* genome. Most of the *C. briggsae* gene set consists of a hybrid set of gene predictions, based on a compilation of results from various gene prediction algorithms. The specific source for each gene prediction is not published yet in WormBase, so it is unclear as to how reliable each gene prediction is at this point in the annotation of the genome. In addition, most of the *C. briggsae* predicted genes are not backed by any experimental expression data (EST's, for example), at least insofar as what is described for each said gene on WormBase. This is an area of continual updating, and it is expected that this particular problem will be resolved in the next several months to a year (Todd Harris, personal communication). Lastly, the *C. briggsae* genome is not yet fully assembled with regards to where each gene is actually located (on which chromosome), so any trends regarding positions of transposable elements (as I completed and described herein with *C. elegans*) is not yet possible to determine via a genomic basis. Again, this area is currently being completed, and expected to be ready soon. Once both the annotation and assembly of the *C. briggsae* genome is more fully complete, I will be able to return to these genes (and other areas of interest) and investigate more fully.

That said, I was able to find out a few things of interest with regards to this third inconclusive subset. Hum-7, for example, exhibited good similarity at both the protein and nucleic acid level with its briggsae ortholog (See Appendix B). It's noteworthy that short disruptions (of a few aa) in otherwise very strong alignment occur at the position of each conserved intron. Both hum-7 and its ortholog are very large genes, thus it was difficult to determine a definitive answer to the conserved intron question, despite my compiling a variety of different alignments (partial genes, with and without section corresponding to the transposable element) – there was no clear indication either way of intron conservation.

The elegans gene mdt-29 presents a case where the elegans gene is much larger than its briggsae ortholog, due to much longer introns across the whole gene. The presence of these much longer introns across the whole gene made assessing the intron conservation using this multiple alignment strategy difficult. One way around this could be to cut out middle portions of each intron in mdt-29, similar to the process I used generally for each element intron, but across the whole gene. In this way the exons would be more likely to align and I would be able to see whether the ortholog did indeed have an intron (albeit very small) in that same region.

The elegans gene srh-291 has good alignment with its briggsae ortholog, but the briggsae annotation shows a predicted gene with predicted introns (as do all briggsae genes at this point). When viewing the alignment, the elegans element

intron in question corresponds to a predicted exon in *briggsae*. At face value this would represent a new intron created by a transposon in *C.elegans*, however, it remains inconclusive at this point due to the present state of the *briggsae* genome annotation. The *elegans* gene *srw-83* presents a similar problem where the *elegans* element intron corresponds to a predicted exon in the *briggsae* ortholog. However, this *briggsae* ortholog does also have an intron in the region, so it seems likely that this is an area of intron conservation and the annotation of there being an exon in this area is incorrect.

The last *elegans* gene in this third group, ZK856.5, has a *briggsae* ortholog that aligns very well to all of the *elegans* gene following intron 3 (where the element is located). It seems likely that if I aligned the upstream flanking sequence to this *briggsae* ortholog with the *elegans* gene, that I might find a similar situation as to that I conducted and is described below for *clec-41*; that being that the surrounding sequence of the intron element in *elegans* (exons 1, 2 and 3 and introns 1 and 2 that did not align with the *briggsae* ortholog) is in an intergenic region in the *briggsae* genome, upstream from the *briggsae* ortholog.

The fourth and final group of *elegans* element intron genes (*clec-41* and T05H4.10) are interesting in that it appears that the transposable element created a new intron when it inserted in each of these genes. The corresponding *briggsae* orthologs for each share a good deal of similarity at the protein and nucleic acid level, and clearly do not contain an intron in the same region.

T05H4.10, seen below in Figures 10, 11 and 12, is a clear example; the Tc9 element forms the entire intron between exons 8 and 9, and the briggsae ortholog does not have an intron. This can clearly be seen in the genomic and cDNA sequences. Additionally, it appears that this may be a region of synteny between the two genomes, and is a region I plan to investigate further in the future.

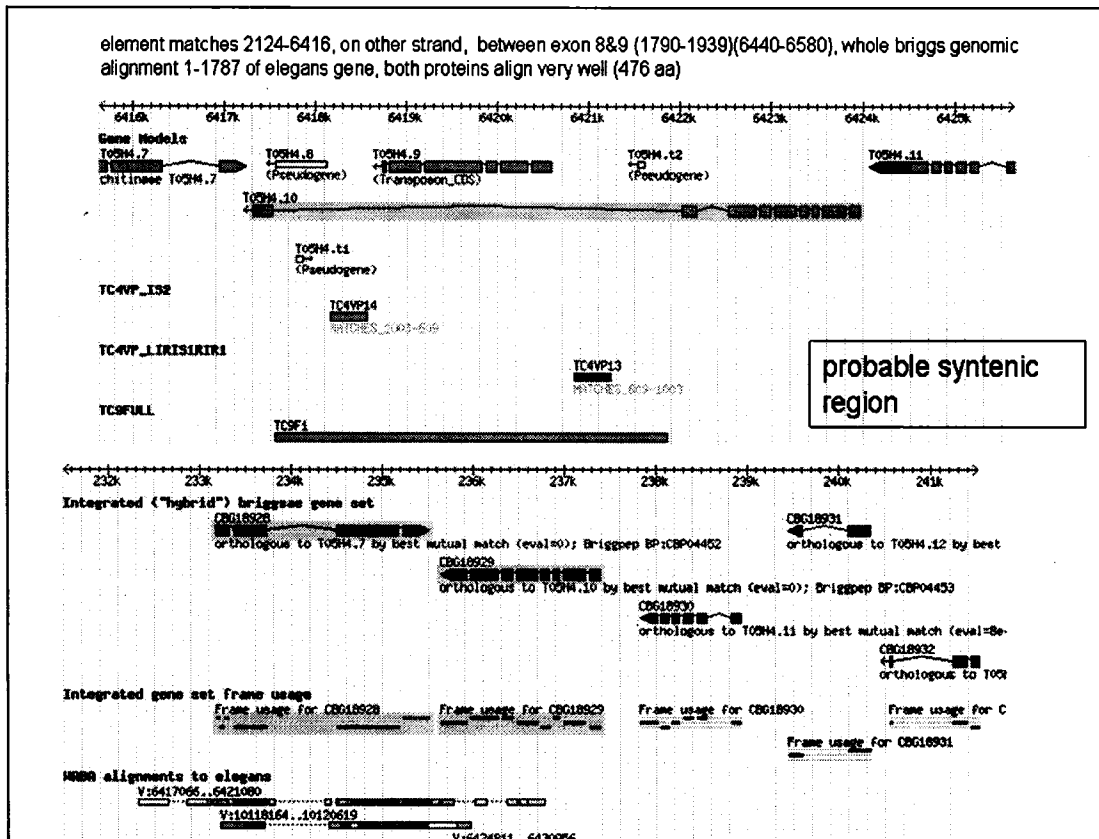


Figure 10 T05H4.10 Part I –element created intron example

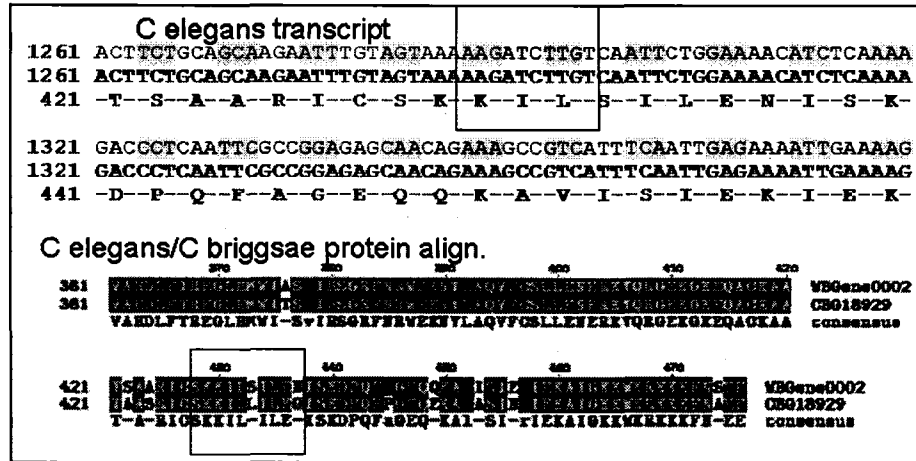


Figure 11 T05H4.10 Part II – element created Intron example

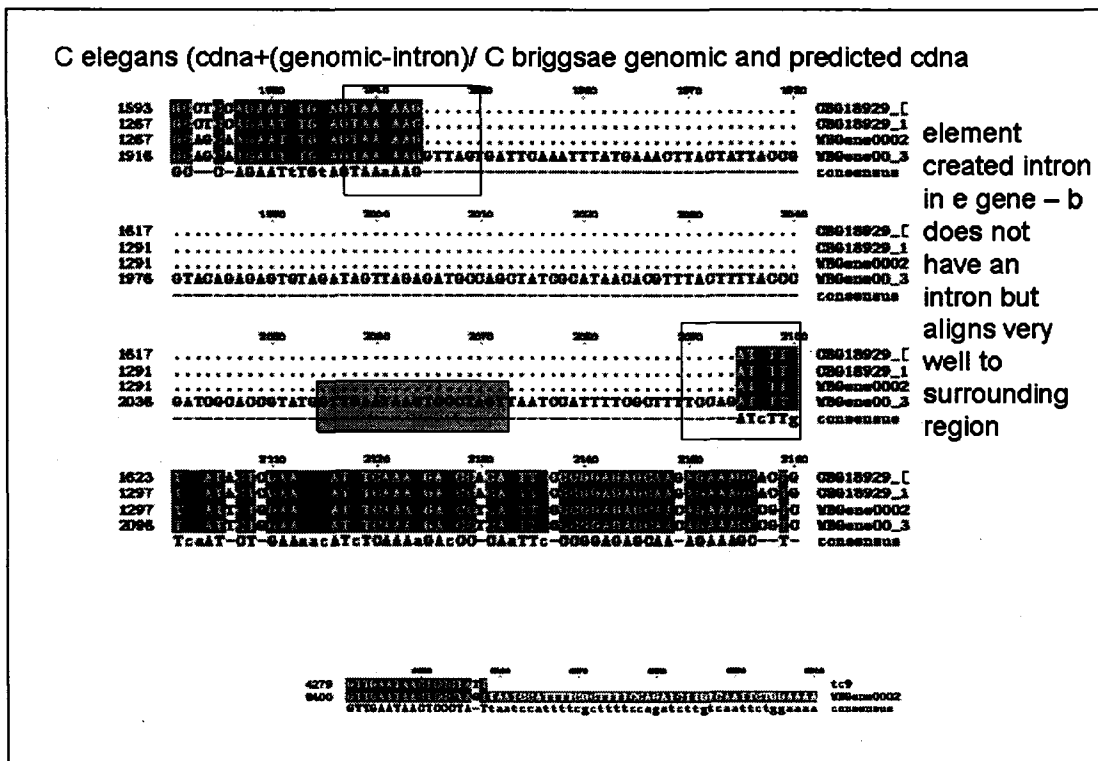


Figure 12 T05H4.10 Part III – element created intron example

A second example of a *C. elegans* intron created by an element is seen in *clec-41* (see Figures 13, 14 and 15 below). This element appears to have inserted after the last coding exon in this gene, and is a region that may be syntenic with *C. briggsae*. The *C. briggsae* ortholog to this gene, CBG09432, aligns very well to all of the *elegans* gene except the end of the gene where the element and non-coding exon reside. I was curious as to what similarity would exist between the flanking sequence of the *briggsae* gene, and discovered that as I expected, the flanking sequence aligns to the last non-coding exon in the *elegans* gene (see Figure 15), and I would suspect to more of this flanking region as well. As with the former example, these two genes and regions are areas of future investigation.

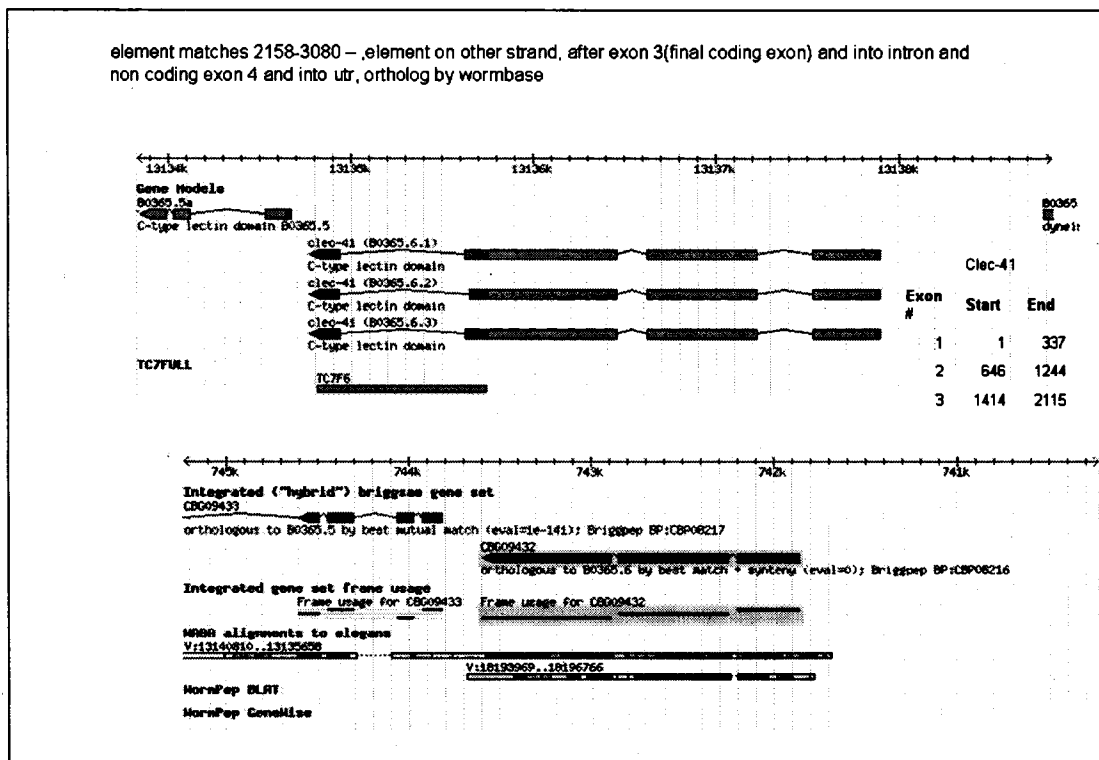


Figure 13 *clec-41* Part I – element created intron example

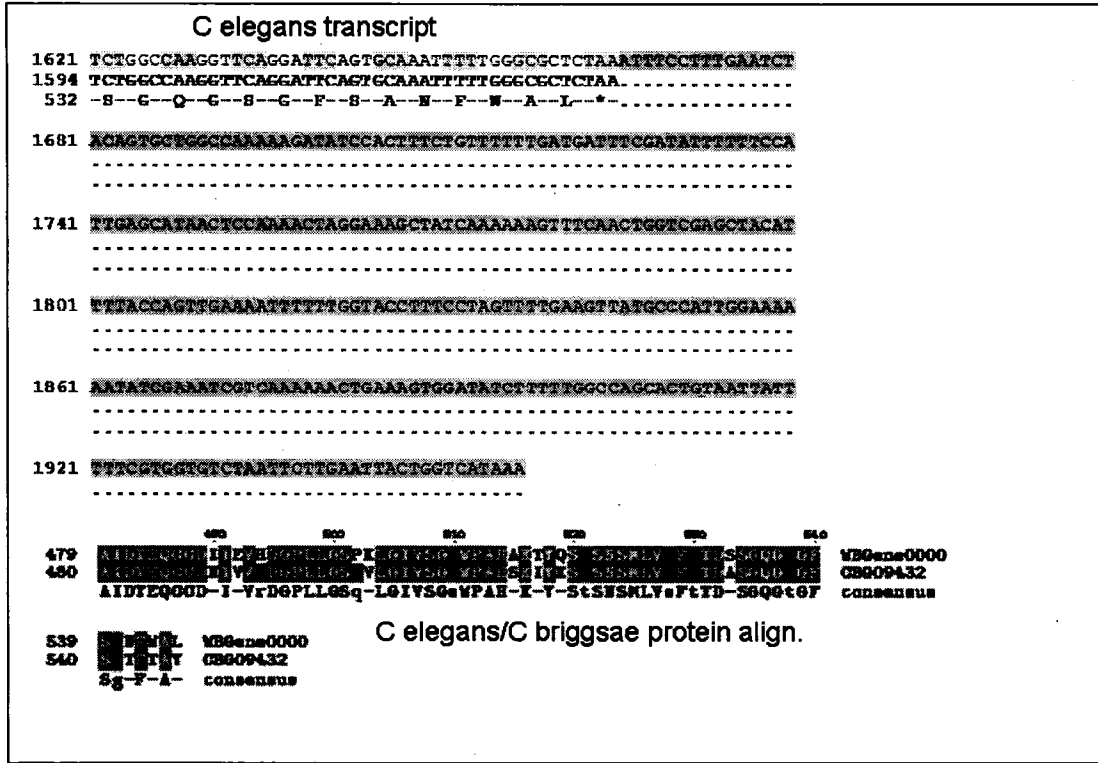


Figure 14 clec-41 Part II – element created intron example

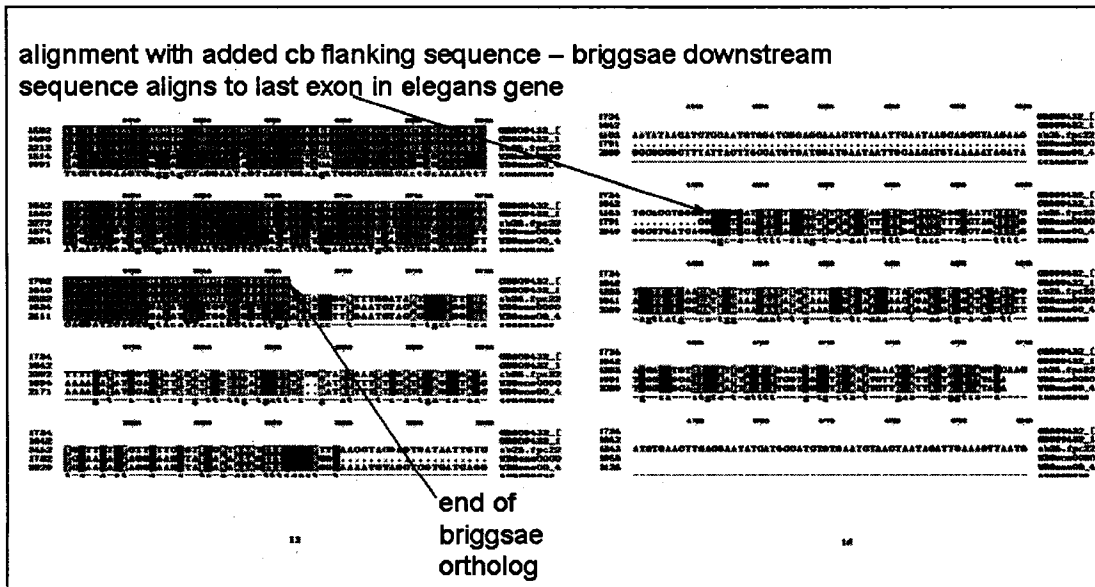


Figure 15 clec-41 Part III – element created intron example

C.briggsae Intron Study (B Intron Study)

As described above, I was interested in determining if any of the transposable elements located within genes encompassed entire introns, and if so, if their ortholog in a related species (*C.elegans*/*C.briggsae* pair) also contained an intron in the same location. I performed this analysis starting with all the *C.elegans* genes for which the transposable element encompassed whole introns (described above), and also beginning with the *C.briggsae* genes of the same (described here). The analyses I performed (various alignments) was essentially the same for the *C.briggsae* intron element genes, with the exception that for the visualization of the transcript and protein alignment for the *C.briggsae* genes, I utilized a different tool, Wise2 (Ensembl, used above for the *C.elegans* transcript/protein alignment view, is not yet available for the *C.briggsae* genome).

Table 12 below displays the results of that investigation. In an effort to remain consistent, I will use the same group descriptors as above in the *elegans* intron study. Of the 11 genes included in this analysis, five had conserved introns in the *C.elegans* orthologs (Group 1). A second group, containing three genes, had no *elegans* ortholog. A third group, containing one gene, is a probable example of a conserved intron, but remains inconclusive at this point. The fourth and final group, containing two genes, are examples (one of which is very evident) of elements creating new introns. Details on all of the above groups and their respective genes are highlighted below.

Four briggsae genes, CBG's 20149, 07789, 24859, and 00653 present clear examples of elements inserting into existing briggsae introns, where the elegans ortholog also has an intron (shorter by approximately the length of the element) in the same region. Briggsae gene CBG 07789 is depicted below in Figures 16 and 17. A fifth member in this group, CBG 09294, is also believed to have a conserved intron, although there are a few bases at the end of the briggsae intron that align over the beginning of the next exon in the elegans gene (see Figure 18), therefore this particular example does not present as clear an answer.

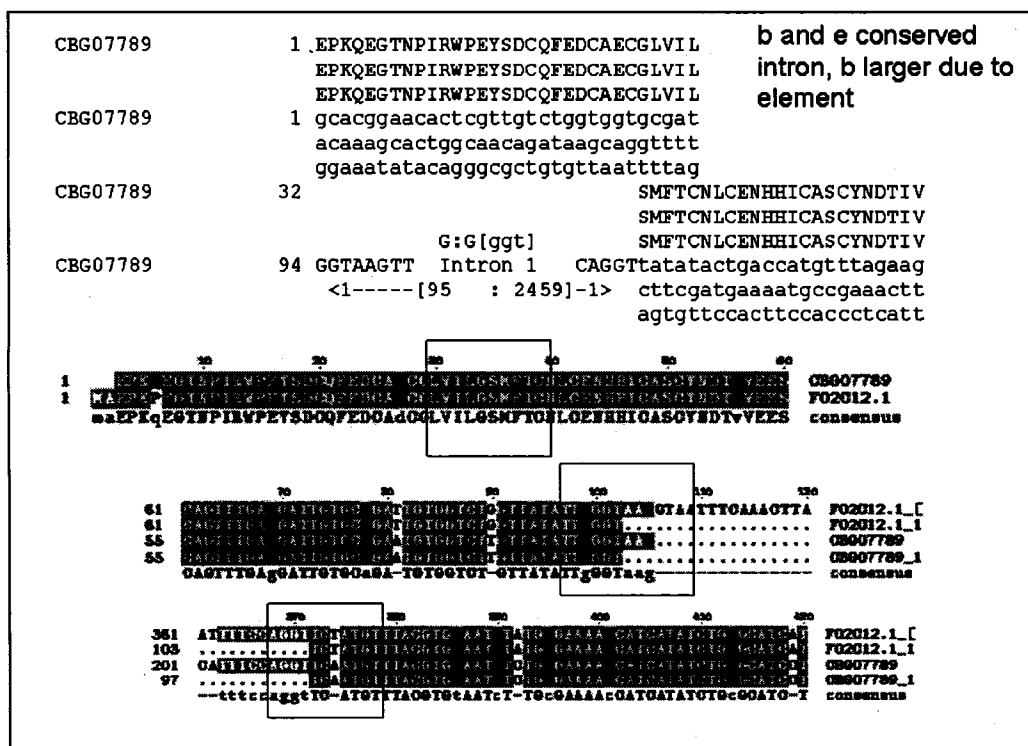


Figure 16 CBG 07789 – Conserved intron example

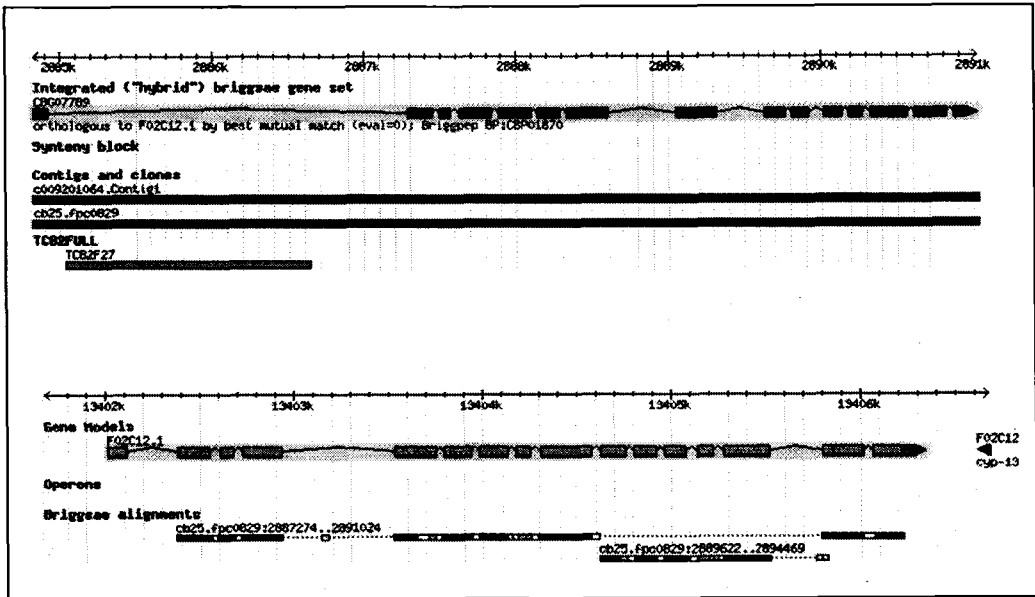


Figure 17 CBG07789/F02C12.1 pair – Conserved intron example

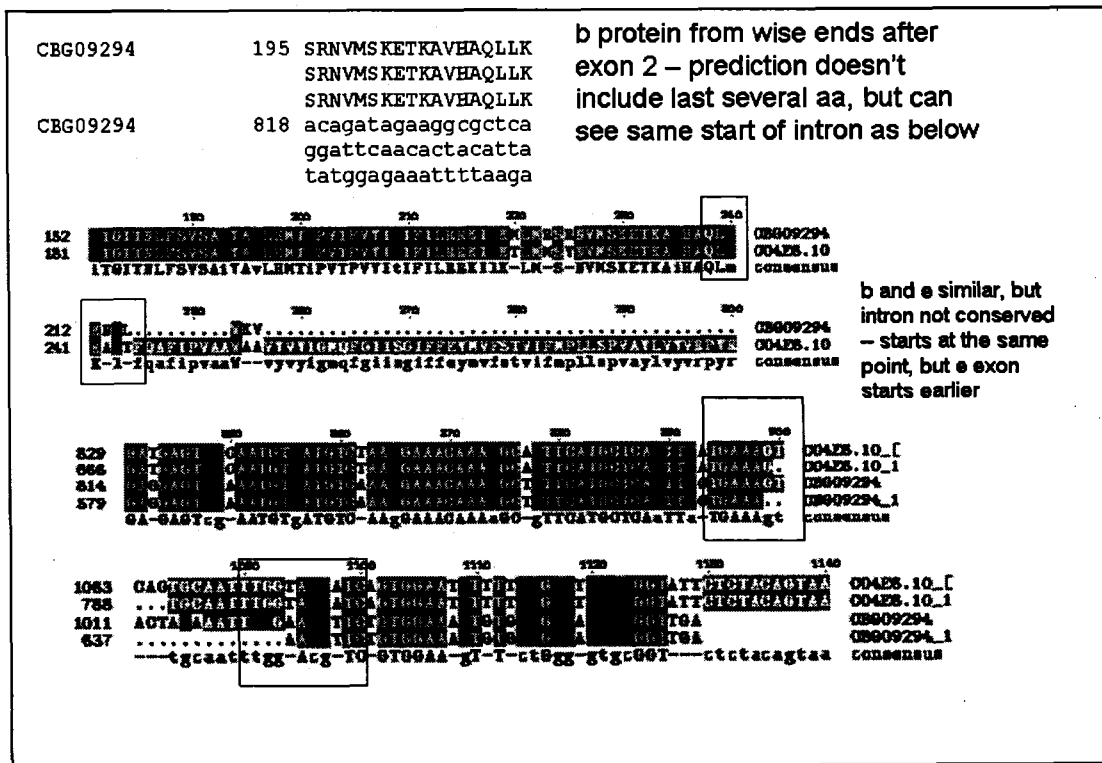


Figure 18 CBG09294 - Conserved intron?

Table 12 *C.briggsae* Intron Study

<i>C.briggsae</i> intron gene/ <i>C.elegans</i> ortholog	Element	b gene/e gene unspliced (bp)	b/e protein (aa)	intron cons v.?	Notes
20149/F43C1.1	B2F40	11019/6144	1041/1036	Y	
07789/F02C12.1	B2F27	6195/4321	828/815	Y	
24859/H06O01.3	B1F6	8480/3729	383/383	Y	
00653/M110.7	B1F8	5770/3848	876/880	Y	
09294/C04E6.10	B2F33	3109/1337	223/337	Y	assume conserved intron based on alignments, exon begins in e gene just a bit before b
05090/W04C9.3	B2F19	9413/4343	488/445	Y?	problems with alignment due to b gene having much larger introns, presume intron conserved
17359/B0252.3c*	B1F13	6177/2103	701/464	N/D**	e gene not orthologous
17587/C06C3.1c*	B2F3	24445/10121	1896/1124	N/D**	e gene not orthologous
10725/na	B2F18	4809	777	N/D**	no e ortholog
06979/K10B4.5	B2F36	2706/1769	305/344	N	intron made element – clearest example of this
20945/C50E10.6	B1F2	3076/2121	346/365	N	b and e similar e has exon where b intron, potentially element made intron, but doesn't encompass full b intron
* indicates no ortholog listed in WormBase, thus BEST BLASTP match listed was used **N/D either no similarity existed between elegans gene or an ortholog was not available, thus question of intron conservation could not be determined					

Three genes were not able to be assessed with regards to intron conservation due to the fact that no elegans ortholog was found. For two of these genes, CBG's 17359 and 17587, the best BLASTP match was utilized as a potential ortholog (same as discussed in Elegans intron study), but did not share any significant similarity with their briggsae counterpart based on the protein and

genomic alignments. For this part of my study, it turned out that these two elegans “orthologs” were the only ones I identified by using the best BLASTP match, all others were listed as orthologs in WormBase. The third member of this no ortholog group, CBG 10725, had neither an elegans ortholog listed, nor an elegans best BLASTP match (although best BLASTP matches in other organisms were available).

For this group, containing CBG 05090, I was unable to determine with certainty whether it had a conserved intron with its elegans ortholog, although there is some evidence that it does. It appears that this may be an example of an issue with the briggsae annotation, as was described in detail earlier. Briefly, in this example, the alignment of both genes show introns that begin in the same location, but the briggsae gene’s next exon begins before the elegans ortholog’s does (see Figure 19). Again, future updating of the briggsae genome should eliminate this ambiguity, as gene predictions would be backed by experimental evidence (such as EST data).

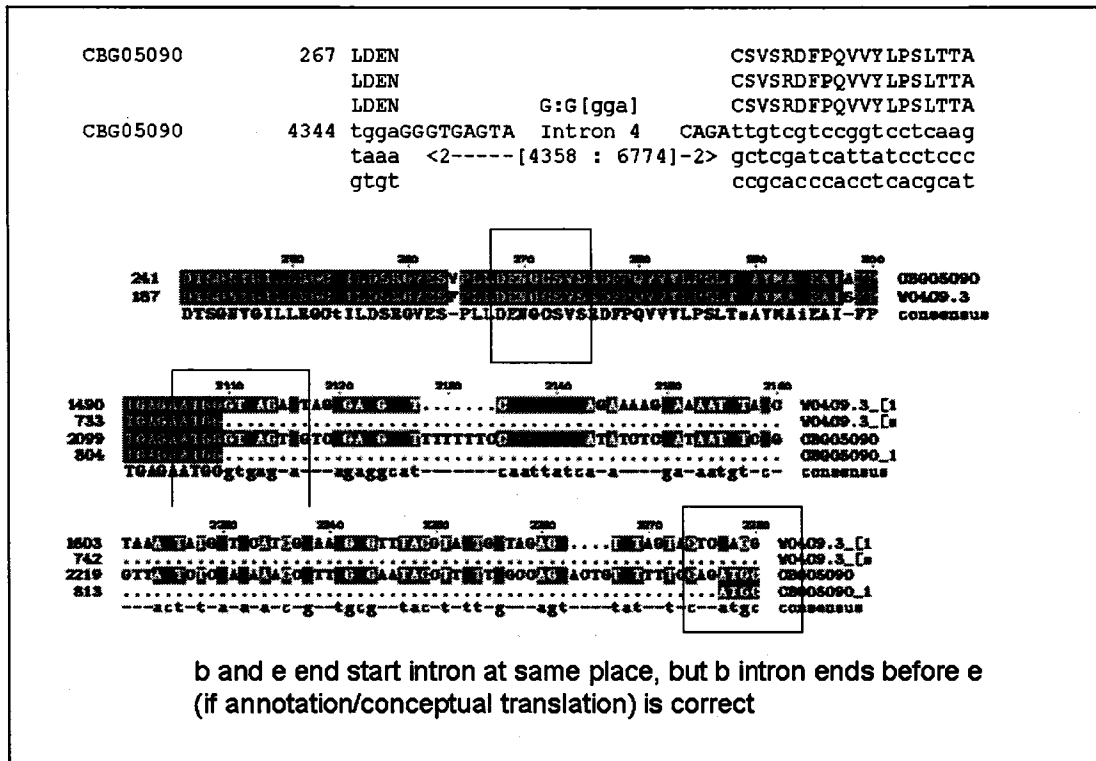


Figure 19 CBG05090 - Conserved intron?

As was also seen in the elegans intron study, I found examples of briggsae genes that appear to have had introns created by transposable elements jumping into an existing exon and splitting it apart. There are two briggsae genes, CBG 06979 and 20945, for which this appears to be the case. CBG 06979 presents the most persuasive evidence for this type of intron creation by a transposable element, as the element itself clearly encompasses the entire intron, and there is excellent similarity in the alignments of the two genes, with most of the element removed from the briggsae gene (see Figures 21, 22, and 23).

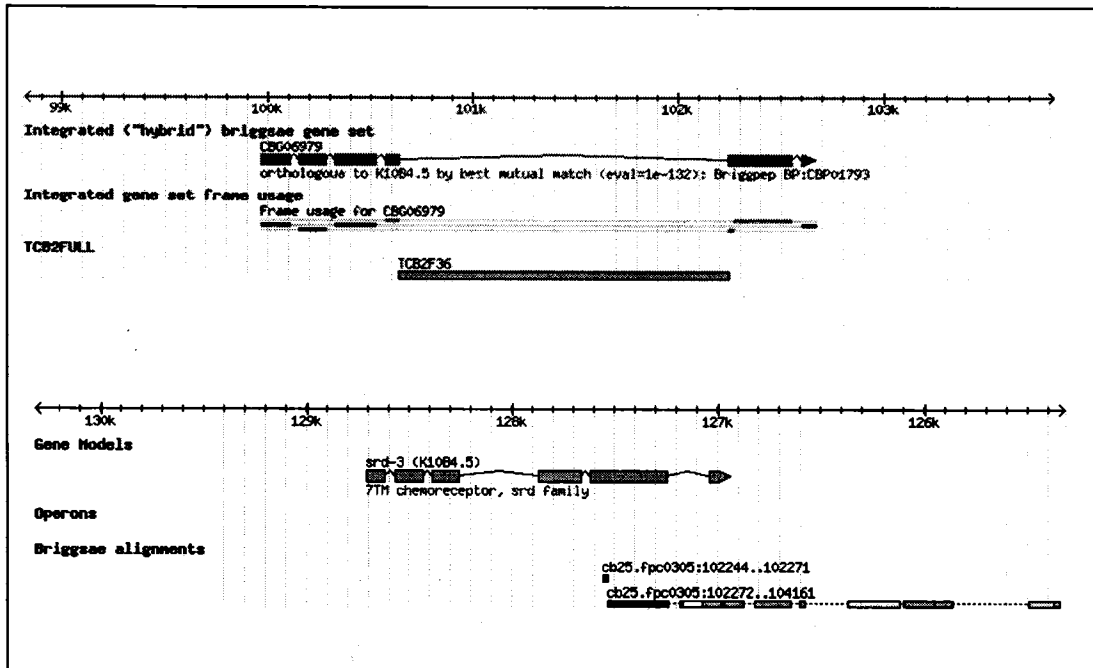


Figure 20 CBG06979 - Element created intron

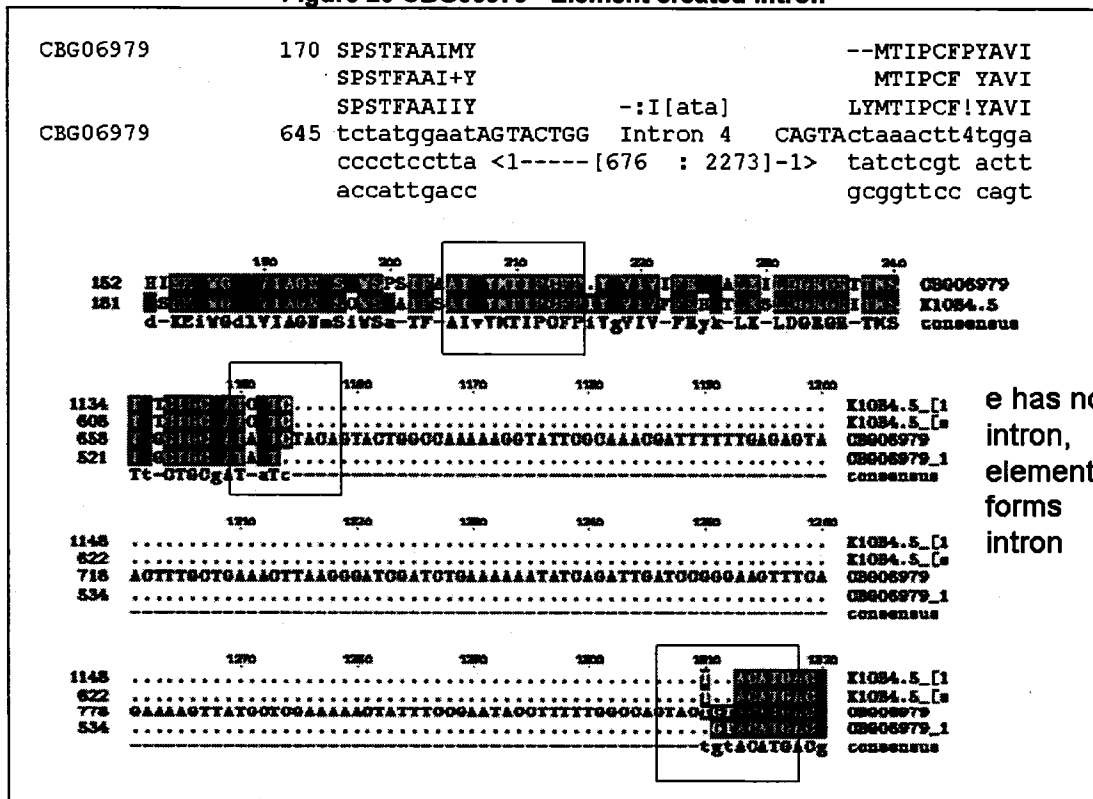


Figure 21 CBG06979 - Element created intron

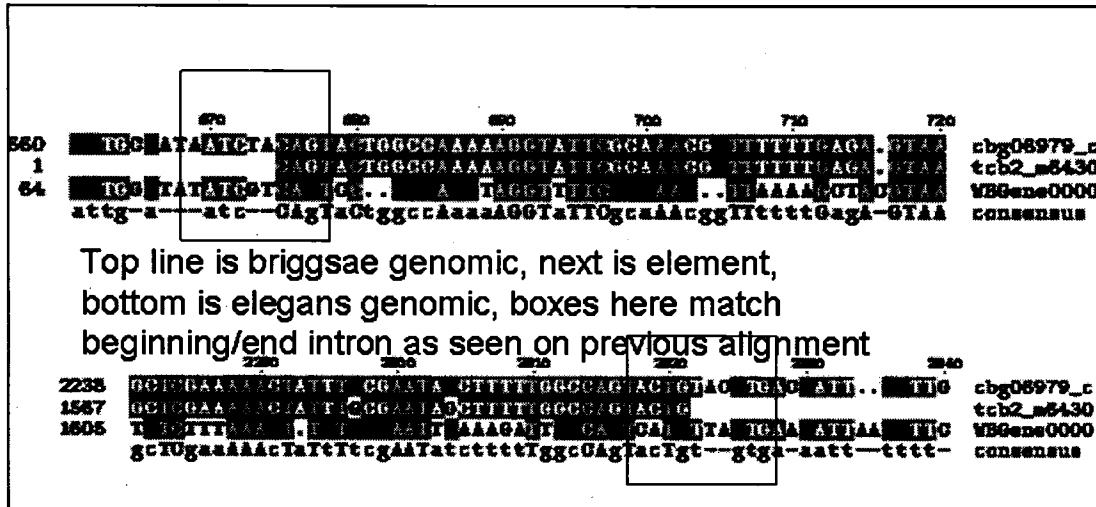


Figure 22 CBG06979 - Element created intron

CBG20945 is a potential gene for which an element created an intron, as there is good similarity in the regions surrounding the element intron, and the elegans ortholog clearly does not have an intron in this region. The only inconclusive part to this particular gene pair is that the element does not appear to encompass the entire briggsae intron (there is extra intronic sequence on either side of the alignment to the element), thus leaving the question of how it could have created an intron but still have extra intron sequence surrounding it. What this suggests is that the inserted element activated cryptic splice sites to create this new intron. This particular gene remains inconclusive at this point. Once the briggsae annotation is more complete, it would be interesting to return to this gene to see if its exon/intron structure had changed.

Additional Analysis

I also wanted to ascertain any information regarding the relative phenotypes of these genes (if known) and any expression patterns (again, if known). My larger research hypothesis regarding transposons potentially having a role in genome stability could be addressed by looking for patterns with regards to these characteristics (i.e. if similar expression patterns could be found between this subset of genes, one could argue that the transposons were aiding in their pattern of expression). Unfortunately, at this time, there does not appear to be an efficient means for locating and analyzing both RNAi phenotypes and global expression patterns, such that every *C.elegans* or *C.briggsae* gene would be included in the analysis. This analysis is something that I plan to pursue in further collaborative work in this area.

Future Directions

While I was unable at this time to determine much information regarding whether any patterning regarding positions of transposable elements are present in other organisms, I was able to isolate several orthologs (other than *C.briggsae*) relative to transposable elements in *C.elegans*. I did perform a search on Ensembl using the list of genes from the *C.elegans* intron study described here to see if there were identified orthologs in other species, and in fact there are. For example, for *C.elegans* gene F30F8.10, there are orthologs in bee, cow, chicken,

chimpanzee, dog, *Drosophila*, Fugu, human, mosquito, mouse, opossum, rhesus, rat, tetraodon, and zebrafish. I intend to continue work in this area, and subsequently analyze each of these orthologs to ascertain whether they have conserved introns as I did with the *C.briggsae* genes described in this project. This will only help to shine further light on the potential role of transposable elements in their respective host genomes.

CHAPTER IV

DISCUSSION

I have presented the results of my bioinformatic analysis of transposons in the *C.elegans* genome. The inspiration for this work has come from a desire to examine the potential functional role for transposable elements in the genomes they inhabit. There has been a longstanding debate over whether these elements are selfish DNA (a classically held belief) or if they may be utilized by the host genome for some functional role. Recently established links between transposons, RNAi, and chromatin-level control of gene expression suggest one possible mechanism to which transposons might play such a role. In this model transposons serve as targets for chromatin modifications mediated by RNAi. These modifications could then potentially regulate nearby genes. In order to test this model, I determined the locations of all the fixed transposable elements in the genome of the model organism, *C.elegans*. I was able to locate 276 elements, consisting of 84 full length and 192 partial elements dispersed throughout the *C.elegans* genome.

I decided to use a $\geq 90\%$ sequence identity for the cutoff for hits in this project, as this seemed to be the best method for both finding an abundant group of elements (both full and partial) as well as being confident that each hit did in fact

correspond to an element (either full or part thereof). Past studies of this kind (i.e. the Fisher et. al. study of 2003 mentioned in results) have used lower similarity percentages, but I did not find any major discrepancies between that study and mine. Specifically, I compiled a chart that correlated this previous data on the full elements in the *C.elegans* genome with my annotated elements described here (see TC_KENICK_FISCHER_COMPARISON_TABLE on supplemental CD). The only discrepancies seen between the Fischer study and mine described herein were distinctions between what constituted a full or partial element. In the Fischer study, since they were only looking for full elements, all of their "hits" of significance were classified as full elements. I found 10 of their full hits that I have classified as partial hits. This distinction is the only difference between their results and mine; as far as locations of full elements in the *C.elegans* genome are concerned, (no published data currently exist for fragments).

That said, since I did use a high cutoff for hits, it is assumed that some elements may have been missed using this approach. I intend in the future to go back and further annotate the elements that constitute BLAST hits of lower significance (80-90% seems like an obvious choice for this next tier), and will annotate these new hits with additional information concerning their relative similarities. As I also mentioned in the results, I do not expect any great variation with regards to global trends in localization of elements, as when I began this study I visualized these hits of lower significance, and their distribution was similar (in that there

were no distinctive patterning on a global scale) to that seen and described for the $\geq 90\%$ hits.

Of the 276 elements that I located in this manner, I found that elements were spread fairly evenly with respect to linkage groups (i.e. I did not find Tc1 located on only one linkage group), as was expected. Additionally, I found elements located across all areas of each linkage group. This was in contrast to reports that found elements located primarily on “gene poor” ends of chromosomes. In fact, my more specific analysis of genome position of these elements revealed that most elements reside in gene “average” regions of chromosomes. Most striking are the elements on Linkage Groups V and X, that (on average) reside in areas of relatively high gene density. These previous studies I have mentioned only correlated locations of transposable elements to gene density on a global scale (i.e. most elements located on the ends of chromosomes, which generally are considered gene poor, in comparison to gene rich centers of chromosomes). Thus, my analysis offers a more specific and local view of the correlations between locations of transposable elements and gene density.

Another interesting feature that was revealed by my analysis was the arrangement of fragmented elements in the *C.elegans* genome. There were very few fragments which retained the transposase encoding region. The majority of the fragmented elements identified consisted of some combination of inverted

repeats (IR's). A major subset of these elements were found as a pair of IR's in the same region (i.e. an LIR and and RIR). Of these IR pairs, some were IR's that were next to each other (with no sequence between them), that were termed overlapping fragments (FRAG) in this investigation. The other main portion of these paired IR's consisted of inverted repeats with sequence between them of unknown origin. Further investigation of these fragments should help elucidate what the sequence between the inverted repeats is, which would further aid in our understanding of the relative stability of transposons in the host genome.

One last item of interest with regard to the locations of elements on a global scale came because of my question regarding locations of elements by linkage group. Linkage Group III had the lowest number of full elements (7), which seems odd, as one would expect that LGX would contain the fewest (and has been assumed in the past). Linkage Group III also contained one of the lowest numbers of fragments (14), along with LGII. Interestingly, LGII contained the highest amount of full elements (23). It is unclear why LGII would contain the highest number of full elements but the lowest number of fragments. Perhaps there is something at work on LGII that helps to preserve these full elements?

To further analyze this question of a functional role for transposons in their host genomes, several future possibilities are obvious. Since the hypothesis is that these elements exhibit some functional role, one might expect that this would result in regulation of gene expression, and thus, you would find clusters of

genes that were expressed similarly located nearby these elements. The best evidence for this would be in analysis of readily available expression data for *C.elegans*. This analysis awaits further organization and analysis as the current state of this expression data is such that it is not directly amenable to producing a clear answer. For example, you can find expression data on a gene in any type of developmental or mutational state, but this same data is not arranged so that you can see it by location within the genome. I am interested in querying and organizing this vast expression database in such a way that you can look at locations within the *C.elegans* genome and find out what the relative levels of expression for each gene in a particular genomic region are. In this way, it would be clear where there were clusters of genes that were expressed in similar ways. Obviously, this would also have to be specific to particular conditions for all the genes in that region (i.e. stages of development). The data currently does not exist in this type of organizational framework, but it only awaits some proper querying and data mining.

Additionally, once such clusters of similar gene expression were found, an overlay of this map and the map already available in WormBase to which my transposable elements have been annotated could be made. In this way, I would be able to quickly identify any possible regulatory clusters to which transposable elements also were located. This step should in fact be quite straightforward, as I already know how to create and upload such expression data (once organized by chromosomal location) directly into the WormBase browser. The final step

would then be to test this hypothesis directly by knocking out the element in this region by genetic means and observing the effect on subsequent expression of genes in this area.

In addition to the expectation of regulatory gene clusters being expressed in a similar manner, you might also expect that these clusters would be conserved across species. One first step in looking at this idea involves a comparative genomics approach, where an analysis of locations of transposable elements between two species is conducted. While I was able to locate transposable elements in the sister species *C.briggsae*, further analysis of potential similar gene clusters (syntenic regions) is not definitive as the *C.briggsae* genome lacks a genetic map. I was able to isolate several areas of probable synteny within the analysis of several elements (*C.elegans* genes T05H4.10 and F30F8.10), but this awaits further analysis and study upon completion of the *C.briggsae* genetic map (currently underway). Additionally, I could search for these regions of synteny in the future for other nematode species that are currently being sequenced.

A final area of interest in this project was regarding the location of elements with respect to genetic position. Previous reports have stated that elements are primarily found in intergenic regions, where they would have the smallest effect on the genome (the transposable element as junk or selfish DNA hypothesis). I found that elements were almost equally likely to be found in gene regions (41%) or between genes (59%). You would expect that if these elements served no

role or function that they would be removed over time by the genomes that they inhabit. Thus, you would expect to find any remaining elements existing solely or largely between genes, and not residing in them.

Of the *C.elegans* and *C.briggsae* genes that contained full element introns that I investigated in this work (15 and 11 respectively – representing those elements that appeared to encompass most or all of an intron), there were several examples in both genomes (5-*C.elegans*, 4-*C.briggsae*) that appear to be elements that inserted into pre-existing introns. There were also a few examples (2 genes in both *C.elegans* and *C.briggsae*) that present evidence of an element actually creating a new intron, but inserting into an exon and splitting it apart. This latter group is quite intriguing, as it points to a potential role for these elements. You would expect that if these elements served no functional role, that novel intron creation would not be permitted by the host genome (it could potentially alter the resultant encoded protein). Furthermore, since there are already published examples of regulatory elements being located within introns of *C.elegans* genes (i.e. *pal-1* as explained in introduction), these elements might also be serving some role. That is not to imply that all of these elements have roles within the genome, but it does suggest that potentially a subset of them (perhaps the elements that create novel introns or insert into pre-existing introns) could serve some regulatory role. Again, further investigation into this, utilizing available expression data as described above, would assist in providing evidence of this.

The results of these intron studies I have reported are admittedly for a subset of these elements located within gene regions, and analyzing the entire set of these is an obvious next step. Additionally, locating and analyzing elements in other nematodes and other taxa will aid in a better understanding of the potential function of the same. I have begun looking for these elements in other taxa briefly (as also detailed in the results) by looking for orthologs of the full elegans elements using available information in Ensembl. I have a small set of genes (5) for which there do appear to be orthologs in diverse species, and these await further investigation.

Since the particular method I wanted to use in this research involved a desire to have a “permanent” record of where each element (full and partial) was located, I have compiled a series of annotation files (located on supplemental CD) for all of the same. I uploaded each file into WormBase while conducting this research, and at this time, these files have not yet been made public. One of my contributions to the scientific community at large will be to have these files published as permanent additions to the WormBase database. This will not only aid my future work in this area (providing the convenience of all annotations permanently available), but should also aid others interested in transposable elements and their patterning (or lack thereof) concerning genomic position. The most novel aspect of this annotation is the fragments I located, as no one to date has published anything regarding positions of transposable element fragments.

The work described in this thesis was exclusively bioinformatic based, and utilization of this type of available analysis is a direction that biology is heading. Massive amounts of sequence data are now available, and for a substantial group of diverse species, expression and other experimental data are now available as well. One of the next major steps in this process of understanding the complex nature of genes and genomes is to connect and network this data in a way that we can efficiently address these complex scientific questions. One area that is now being compiled is connecting this vast amount of expression data with genomic location, as I mentioned as one of the next steps in my work. In the near future, you should be able to quickly identify clusters of genes that are expressed in similar ways by choosing a region of the genome of interest by bioinformatic means. These types of analyses and networking of complex data sets will undoubtedly drastically change the way and speed in which we can help to answer questions of global gene expression and patterns therein.

REFERENCES

- Agrawal N, Dasaradhi PV, Mohammed A, Malhotra P, Bhatnagar RK, Mukherjee SK. (2003). RNA interference: biology, mechanism, and applications. *Microbiol Mol Biol Rev.* 67(4), 657-85.
- Alfonso, A., Grundahl, K., McManus, J.R., Asbury, J.M., Rand, and J.B. (1994). Alternative splicing leads to two cholinergic proteins in *C.elegans*. *J. Mol. Biol.* 24, 627–630.
- Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (1990) Basic local alignment search tool. *J. Mol. Biol.* 215:403-10.
- Ambros, V., Lee, R.C., Lavanway, A., Williams, P.T., and Jewell, D. (2003). MicroRNAs and other tiny endogenous RNAs in *C.elegans*. *Curr. Biol.* 13, 807–818.
- Ambros, V. (2001) microRNAs: tiny regulators with great potential. *Cell* 107, 823–826.
- Anderson, P. (1995). Mutagenesis. *Methods Cell. Biol.* 48, 31–58.
- Barnes, T.M., Kohara, Y., Coulson, A., and Hekimi, S. (1995). Meiotic recombination, noncoding DNA and genomic organization in *C.elegans*. *Genetics.* 141, 159–179.
- Barrett, P.L., Fleming, J.T., and Gobel, V. (2004). Targeted gene alteration in *C.elegans* by gene conversion. *Nat. Genet* 36, 1231–1237.
- Berezikov, E., Bargmann, C.I., and Plasterk, R.H. (2004). Homologous gene targeting in *C.elegans* by biolistic transformation. *Nucleic Acids Res.* 32, e40.
- Berg, D.E., and Howe, M.M. (1989). *Mobile DNA* (Washington, D.C.: American Society for Microbiology).
- Bessereau, J.L., Wright, A., Williams, D.C., Schuske, K., Davis, M.W., and Jorgensen, E.M. (2001). Mobilization of a *Drosophila* transposon in the *C.elegans* germ line. *Nature* 413, 70–74.
- Betran, E., and Long, M. (2002). Expansion of genome coding regions by acquisition of new genes. *Genetica* 115, 65–80.

Bessereau, J.-L. Transposons in *C.elegans* (January 18, 2006), WormBook, ed. The *C.elegans* Research Community, WormBook, doi/10.1895/wormbook.1.70.1, <http://www.wormbook.org>.

Birney, E., Clamp, M., and Durbin, R., (2004). GeneWise and Genomewise. *Genome Research* 14, 988-995.

Blumenthal, T., and Steward, K. (1997). RNA processing and gene structure. In: *C.elegans* II, D.L. Riddle, T. Blumenthal, B.J. Meyer, J.R. Priess, Eds (Plainview, New York: Cold Spring Harbor Laboratory Press), pp. 117–145.

Blumenthal, T., Evans, D., Link, C.D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W.L., Duke, K., Kiraly, M., et al. (2002). A global analysis of *C.elegans* operons. *Nature* 417, 851–854.

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

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

Brennecke, J., Hipfner, D.R., Stark, A., Russell, R.B., and Cohen, S.M. (2003). Bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene *hid* in *Drosophila*. *Cell* 113, 25–36.

Brenner, S. (2000). The end of the beginning. *Science* 287, 2173–2174.

Britten, R.J. (1995). Active gypsy/Ty3 retrotransposons or retroviruses in *C.elegans*. *Proc. Natl. Acad. Sci. USA* 92, 599–601.

Brookfield, J.F. (2005). The ecology of the genome - mobile DNA elements and their hosts. *Nat. Rev. Genet.* 6, 128–136.

Broverman, S., MacMorris, M., and Blumenthal, T. (1993). Alteration of *C.elegans* gene expression by targeted transformation. *Proc. Natl. Acad. Sci. USA* 90, 4359–4363.

Brownlie, J.C., and Whyard, S. (2004). CemaT1 is an active transposon within the *C.elegans* genome. *Gene* 338, 55–64.

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

Carmell, M.A., and Hannon, G.J. (2004). RNase III enzymes and the initiation of gene silencing. *Nat. Struct. Mol. Biol.* 11, 214–218.

- Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V., and Kondrashov, F.A. (2002). Selection for short introns in highly expressed genes. *Nat Genet.* 31, 415–418.
- Chen, C.-C.-G., Simard, M.J., Tabara, H., Brownell, D.R., McCollough, J.A., and Mello, C.C. (2005). A member of the polymerase beta nucleotidyltransferase superfamily is required for RNA interference in *C.elegans*. *Curr. Biol.* in press.
- Claudianos, C., Brownlie, J., Russell, R., Oakeshott, J., and Whyard, S. (2002). maT--a clade of transposons intermediate between mariner and Tc1. *Mol. Biol. Evol.* 19, 2101–2109.
- Coghlan, A., and Wolfe, K.H. (2002). Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. *Genome Res.* 12, 857–867.
- Cogoni, C., and Macino, G. (1999). Posttranscriptional gene silencing in *Neurospora* by a RecQ DNA helicase. *Science* 286, 2342–2344.
- Collins, J., Forbes, E., and Anderson, P. (1989). The Tc3 family of transposable genetic elements in *C.elegans*. *Genetics* 121, 47–55.
- Collins, J., Saari, B., and Anderson, P. (1987). Activation of a transposable element in the germ line but not the soma of *C.elegans*. *Nature* 328, 726–728.
- Collins, J.J., and Anderson, P. (1994). The Tc5 family of transposable elements in *C.elegans*. *Genetics* 137, 771–781.
- Colloms, S.D., van Luenen, H.G., and Plasterk, R.H. (1994). DNA binding activities of the *C.elegans* Tc3 transposase. *Nucleic Acids Res.* 22, 5548–5554.
- Cullen, B.R. (2004). Transcription and processing of human microRNA precursors. *Mol. Cell* 16, 861–865.
- Cutter, A.D., and Payseur, B.A. (2003). Rates of deleterious mutation and the evolution of sex in *Caenorhabditis*. *J. Evol. Biol.* 16, 812–822.
- Dreyfus, D.H., and Emmons, S.W. (1991). A transposon-related palindromic repetitive sequence from *C.elegans*. *Nucleic Acids Res.* 19, 1871–1877.
- Dreyfus, D.H., and Gelfand, E.W. (1999). Comparative analysis of invertebrate Tc6 sequences that resemble the vertebrate V(D)J recombination signal sequences (RSS). *Mol. Immunol.* 36, 481–488.
- Duret, L., Marais, G., and Biemont, C. (2000). Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *C.elegans*. *Genetics* 156, 1661–1669.

- Egilmez, N.K., Ebert, R.H., II, and Shmookler Reis, R.J. (1995). Strain evolution in *C.elegans*: transposable elements as markers of interstrain evolutionary history. *J. Mol. Evol.* 40, 372–381.
- Eide, D., and Anderson, P. (1985). Transposition of Tc1 in the nematode *C.elegans*. *Proc. Natl. Acad. Sci. USA* 82, 1756–1760.
- Eide, D., and Anderson, P. (1988). Insertion and excision of *Caenorhabditis elegans* transposable element Tc1. *Mol. Cell Biol.* 8, 737–746.
- Ellis, R.E., Sulston, J.E., and Coulson, A.R. (1986). The rDNA of *C.elegans*: sequence and structure. *Nucleic Acids Res.* 11, 2345–2364.
- Emmons, S.W., and Yesner, L. (1984). High-frequency excision of transposable element Tc 1 in the nematode *Caenorhabditis elegans* is limited to somatic cells. *Cell* 36, 599–605.
- Emmons, S.W., Yesner, L., Ruan, K.S., and Katzenberg, D. (1983). Evidence for a transposon in *Caenorhabditis elegans*. *Cell* 32, 55–65.
- Fedoroff, N. (1989). Maize Transposable Elements. In *Mobile DNA*, D.E. Berg and M.M. Howe, eds. (Washington, D.C.: American Society for Microbiology), pp. 375–412.
- Finnegan, D.J. (1989). Eukaryotic transposable elements and genome evolution. *Trends in Genetics* 5, 103–107.
- Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E. and Mello, C.C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391,806–811.
- Fischer, S.E., van Luenen, H.G., and Plasterk, R.H. (1999). Cis requirements for transposition of Tc1-like transposons in *C.elegans*. *Mol. Gen. Genet* 262, 268–274.
- Fischer, S.E., Wienholds, E., and Plasterk, R.H. (2003). Continuous exchange of sequence information between dispersed Tc1 transposons in the *Caenorhabditis elegans* genome. *Genetics* 164, 127–134.
- Frame, I.G., Cutfield, J.F., and Poulter, R.T. (2001). New BEL-like LTR-retrotransposons in *Fugu rubripes*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. *Gene* 263, 219–230.
- Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.J., Smirnov, S., Purcell, S., Rehman, B., Elkins, T., Engels, R., Wang, S., Nielsen, C.B., Butler, J., Endrizzi, M., Qui, D., Ianakiev, P., Bell-Pedersen, D., Nelson, M.A., Werner-Washburne, M., Selitrennikoff, C.P., Kinsey, J.A., Braun, E.L., Zelter, A., Schulte, U., Kothe, G.O., Jedd, G., Mewes, W.,

Staben, C., Marcotte, E., Greenberg, D., Roy, A., Foley, K., Naylor, J., Stange-Thomann, N., Barrett, R., Gnerre, S., Kamal, M., Kamvysselis, M., Mauceli, E., Bielke, C., Rudd, S., Frishman, D., Krystofova, S., Rasmussen, C., Metzenberg, R.L., Perkins, D.D., Kroken, S., Cogoni, C., Macino, G., Catcheside, D., Li, W., Pratt, R.J., Osmani, S.A., DeSouza, C.P., Glass, L., Orbach, M.J., Berglund, J.A., Voelker, R., Yarden, O., Plamann, M., Seiler, S., Dunlap, J., Radford, A., Aramayo, R., Natvig, D.O., Alex, L.A., Mannhaupt, G., Ebbole, D.J., Freitag, M., Paulsen, I., Sachs, M.S., Lander, E.S., Nusbaum, C., and Birren, B. (2003). The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* 422, 859–868.

Ganko, E.W., Bhattacharjee, V., Schliekelman, P., and McDonald, J.F. (2003). Evidence for the contribution of LTR retrotransposons to *C.elegans* gene evolution. *Mol. Biol. Evol.* 20, 1925–1931.

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

Garfinkel, D.J., Boeke, J.D., and Locate, G.R. (1985). Ty element transposition: reverse transcriptase and virus-like particles. *Cell* 42, 507-17.

Granger, L., Martin, E., and Segalat, L. (2004). Mos as a tool for genome-wide insertional mutagenesis in *Caenorhabditis elegans*: results of a pilot study. *Nucleic Acids Res.* 32, e117.

Grishok, A., Tabara, H., and Mello, C.C. (2000) Genetic requirements for inheritance of RNAi in *C.elegans*. *Science* 287, 2494–2497.

Grunstein, M. (1997). Histone acetylation in chromatin structure and transcription. *Nature* 389, 349–352.

Guo, S. and Kemphues, K.J. (1995) *par-1*, a gene required for establishing polarity in *C.elegans* embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. *Cell* 81, 611–620.

Gutierrez, A., and Sommer, R.J. (2004). Evolution of *dnmt-2* and *mbd-2*-like genes in the free-living nematodes *Pristionchus pacificus*, *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Nucleic Acids Res.* 32, 6388–6396.

Haber, J.E. (2000). Partners and pathways repairing a double-strand break. *Trends Genet.* 16, 259–264.

Hampsey, M., and Reinberg, D. (2003). Tails of intrigue: phosphorylation of RNA polymerase II mediates histone methylation. *Cell* 113, 429–432.

Hannon, G. (2002). RNA interference. *Nature* 418, 244–251.

- Haren, L., Ton-Hoang, B., and Chandler, M. (1999). Integrating DNA: transposases and retroviral integrases. *Annu. Rev. Microbiol.* 53, 245–281.
- Harris LJ, Prasad S, Rose AM. (1990). Isolation and sequence analysis of *Caenorhabditis briggsae* repetitive elements related to the *Caenorhabditis elegans* transposon Tc1. *J Mol Evol.* Apr;30(4):359-69.
- Harris LJ, Baillie DL, Rose AM. (1988). Sequence identity between an inverted repeat family of transposable elements in *Drosophila* and *Caenorhabditis*. *Nucleic Acids Res.* Jul 11;16(13):5991-8.
- Harris, T.W., Chen, N., Cunningham, F., Tello-Ruiz, M., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bram, K., and Chan, J. (2004). WormBase: a multi-species resource for nematode biology and genomics. *Nucleic Acids Res.* 32, D411–D417.
- Harrison, P.M., Echols, N., and Gerstein, M.B. (2001). Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. *Nucleic Acids Res.* 29, 818–830.
- He, L. and Hannon, G.J. (2004). MicroRNAs: Small RNAs with a big role in gene regulation. *Nat. Rev. Genet.* 5, 522–531.
- Heschl, M.F., and Baillie, D.L. (1989). Identification of a heat-shock pseudogene from *Caenorhabditis elegans*. *Genome* 32, 190–195.
- Higgins D., Thompson J., Gibson T., Thompson J.D., Higgins D.G., Gibson T.J.(1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice. *Nucleic Acids Res.* 22:4673-4680.
- Hodgkin, J. A., and Brenner, S. (1977). Mutations causing transformation of sexual phenotype in the nematode *Caenorhabditis elegans*. *Genetics* 123, 301-13.
- Izsvak, Z., Ivics, Z., and Plasterk, R.H. (2000). Sleeping Beauty, a wide host-range transposon vector for genetic transformation in vertebrates. *J. Mol. Biol.* 302, 93–102.
- Jacobson, J.W., Medhora, M.M., and Hartl, D.L. (1986). Molecular structure of a somatically unstable transposable element in *Drosophila*. *Proc. Natl. Acad. Sci. USA* 83, 8684–8688.
- Jansen, G., Hazendonk, E., Thijssen, K.L., and Plasterk, R.H. (1997). Reverse genetics by chemical mutagenesis in *Caenorhabditis elegans*. *Nat. Genet.* 17, 119–121.

Johnston, R.J., and Hobert, O. (2003). A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* 426, 845–849.

Juarez, M.T., Kui, J.S., Thomas, J., Heller, B.A., and Timmermans, M.C. (2004). microRNA-mediated repression of rolled leaf1 specifies maize leaf polarity. *Nature* 428, 84–88.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. (2003). Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 421, 231–237.

Kazazian, H.H., Jr., (2004). Mobile elements: drivers of genome evolution. *Science* 303, 1626–1632.

Kelly, W.G., and Fire, A. (1998). Chromatin silencing and the maintenance of a functional germline in *Caenorhabditis elegans*. *Development* 125, 2451–56.

Kelly, W.G., Xu, S., Montgomery, M.K., and Fire, A. (1997). Distinct requirements for somatic and germline expression of a generally expressed *Caenorhabditis elegans* gene. *Genetics* 146, 227–38.

Ketting, R.F., Fischer, S.E., and Plasterk, R.H. (1997). Target choice determinants of the Tc1 transposon of *Caenorhabditis elegans*. *Nucleic Acids Res.* 25, 4041–47.

Ketting, R.F., Haverkamp, T.H., van Luenen, H.G., and Plasterk, R.H. (1999). Mut-7 of *C.elegans*, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. *Cell* 99, 133–141.

Kidner, C.A. and Martienssen, R.A. (2004). Spatially restricted microRNA directs leaf polarity through ARGONAUTE1. *Nature* 428: 81–84.

Kiff, J.E., Moerman, D.G., Schriefer, L.A., and Waterston, R.H. (1988). Transposon-induced deletions in *unc-22* of *C.elegans* associated with almost normal gene activity. *Nature* 331, 631–633.

Korf, I., Flicek, P., Duan, D., and Brent, M.R. (2001). Integrating genomic homology into gene structure prediction. *Bioinformatics* 17(Suppl 1), S140–S148.

Kouzarides, T. (2002). Histone methylation in transcriptional control. *Curr. Opin. Genet. Dev.* 12, 198–209.

Lampe, D.J., Churchill, M.E., and Robertson, H.M. (1996). A purified mariner transposase is sufficient to mediate transposition in vitro. *EMBO J.* 15, 5470–79.

- Lampe, D.J., Grant, T.E., and Robertson, H.M. (1998). Factors affecting transposition of the Himar1 mariner transposon in vitro. *Genetics* 149, 179–187.
- Le, Q.H., Turcotte, K., and Bureau, T. (2001). Tc8, a Tourist-like transposon in *Caenorhabditis elegans*. *Genetics* 158, 1081–88.
- Levitt, A., and Emmons, S.W. (1989). The Tc2 transposon in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* 86, 3232–36.
- Li, W., and Shaw, J.E. (1993). A variant Tc4 transposable element in the nematode *C.elegans* could encode a novel protein. *Nucleic Acids Res.* 21, 59–67.
- Liao, L.W., Rosenzweig, B., and Hirsh, D. (1983). Analysis of a transposable element in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* 80, 3585–89.
- Lippman, Z., Gendrel, A.V., Black, M., Vaughn, M.W., Dedhia, N., McCombie, W.R., Lavine, K., Mittal, V., May, B., Kasschau, K.D., et al. (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature* 430, 471–476.
- Liu, J., Carmell M.A., Rivas, F.V., Marsden C.G., Thomson, J.M., Song, J.J., Hammond, S.M., Joshua-Tor, L., and Hannon, G.J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. *Science* 305, 1437–1441.
- Loftus, B.J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I.J., Fraser, J.A., Allen, J.E., Bosdet, I.E., Brent, M.R., Chiu, R., Doering, T.L., Donlin, M.J., D'Souza, C.A., Fox, D.S., Grinberg, V., Fu, J., Fukushima, M., Haas, B.J., Huang, J.C., Janbon, G., Jones, S.J., Koo, H.L., Krzywinski, M.I., Kwon-Chung, J.K., Lengeler, K.B., Maiti, R., Marra, M.A., Marra, R.E., Mathewson, C.A., Mitchell, T.G., Pertea, M., Riggs, F.R., Salzberg, S.L., Schein, J.E., Shvartsbeyn, A., Shin, H., Shumway, M., Specht, C.A., Suh, B.B., Tenney, A., Utterback, T.R., Wickes, B.L., Wortman, J.R., Wye, N.H., Kronstad, J.W., Lodge, J.K., Heitman, J., Davis, R.W., Fraser, C.M., and Hyman, R.W. (2005). The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. *Science* 307, 1321–1324.
- Lohe, A.R., and Hartl, D.L. (2002). Efficient mobilization of mariner in vivo requires multiple internal sequences. *Genetics* 160, 519–526.
- Lohe, A.R., Timmons, C., Beerman, I., Lozovskaya, E.R., and Hartl, D.L. (2000). Self-inflicted wounds, template-directed gap repair, and a recombination hotspot. Effects of the mariner transposase. *Genetics* 154, 647–656.
- Long, M. (2001). Evolution of novel genes. *Curr. Opin. Genet. Dev.* 11, 673–680.

- Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964.
- MacMorris, M.A., Zorio, D.A., and Blumenthal, T. (1999). An exon that prevents transport of a mature mRNA. *Proc. Natl. Acad. Sci. USA* 96, 3813–3818.
- Malik, H.S., and Eickbush, T.H. (2000). NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from *Caenorhabditis elegans*. *Genetics* 154, 193–203.
- Marin, I., Plata-Rengifo, P., Labrador, M., and Fontdevila, A. (1998). Evolutionary relationships among the members of an ancient class of non-LTR retrotransposons found in the nematode *Caenorhabditis elegans*. *Mol. Biol. Evol.* 15, 1390–1402.
- Martin, E., Laloux, H., Couette, G., Alvarez, T., Bessou, C., Hauser, O., Sookhareea, S., Labouesse, M., and Segalat, L. (2002). Identification of 1088 new transposon insertions of *Caenorhabditis elegans*: a pilot study toward large-scale screens. *Genetics* 162, 521–524.
- Matzke, M.A., Mette, M.F., and Matzke, A.J. (2000). Transgene silencing by the host genome defense: Implications for the evolution of epigenetic control mechanisms in plants and vertebrates. *Plant Mol. Biol.* 43, 401–415.
- Meister, G. and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. *Nature* 431: 343–349.
- Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., and Tuschl T. (2004). Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol. Cell* 15, 185–197.
- Mori, I., Moerman, D.G., and Waterston, R.H. (1988). Analysis of a mutator activity necessary for germline transposition and excision of Tc1 transposable elements in *Caenorhabditis elegans*. *Genetics* 120, 397–407.
- Motamedi, M.R., Verdel, A., Colmenares, S.U., Gerber, S.A., Gygi, S.P., and Moazed, D. (2004). Two RNAi complexes, RITS and RDRC, physically interact and localize to noncoding centromeric RNAs. *Cell* 119, 789–802.
- Mounsey, A., Bauer, P., and Hope, I.A. (2002). Evidence suggesting that a fifth of annotated *Caenorhabditis .elegans* genes may be pseudogenes. *Genome Res.* 12, 770–775.
- Mourrain, P., Beclin, C., Elmayan, T., Feuerbach, F., Godon, C., Morel, J.B., Jouette, D., Lacombe, A.M., Nikic, S., Picault, N., et al. (2000). Arabidopsis SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance. *Cell* 101, 533–542.

Noma, K., Sugiyama, T., Cam, H., Verdel, A., Zofall, M., Jia, S., Moazed, D., and Grewal, S.I. (2004). RITS acts in cis to promote RNA interference-mediated transcriptional and posttranscriptional silencing. *Nat. Genet.* 36, 1174–1180.

Nowrousian, M., Wurtz, C., Poggeler, S., and Kuck, U. (2004). Comparative sequence analysis of *Sordaria macrospora* and *Neurospora crassa* as a means to improve genome annotation. *Fungal Genet. Biol.* 41, 285–292.

Oosumi, T., Garlick, B., and Belknap, W.R. (1995). Identification and characterization of putative transposable DNA elements in solanaceous plants and *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* 92, 8886–8890.

Oosumi, T., Garlick, B., and Belknap, W.R. (1996). Identification of putative nonautonomous transposable elements associated with several transposon families in *Caenorhabditis elegans*. *J. Mol. Evol.* 43, 11–18.

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

Palatnik, J.F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J.C., and Weigel, D. (2003). Control of leaf morphogenesis by microRNAs. *Nature* 425, 257–263.

Parkinson, J., Mitreva, M., Whitton, C., Thomson, M., Daub, J., Martin, J., Schmid, R., Hall, N., Barrell, B., Waterston, R.H., et al. (2004). A transcriptomic analysis of the phylum Nematoda. *Nat. Genet.* 36, 1259–1267.

Peterson, C.L., and Laniel, M.A. (2004). Histones and histone modifications. *Curr. Biol.* 14, R546–R551.

Pfeffer, S., Zavolan, M., Grasser, F.A., Chien, M., Russo, J.J., Ju, J., John, B., Enright, A.J., Marks, D., Sander, C., et al. 2004. Identification of virus-encoded microRNAs. *Science* 304, 734–6.

Plasterk, R.H. (1991). The origin of footprints of the Tc1 transposon of *Caenorhabditis elegans*. *EMBO J.* 10, 1919–25.

Plasterk, R.H., and Groenen, J.T. (1992). Targeted alterations of the *Caenorhabditis elegans* genome by transgene instructed DNA double strand break repair following Tc1 excision. *EMBO J.* 11, 287–290.

Plasterk, R.H., Izsvak, Z., and Ivics, Z. (1999). Resident aliens: the Tc1/mariner superfamily of transposable elements. *Trends Genet.* 15, 326–332.

Plasterk, R.H.A., and van Luenen, H.G.A.M. (1997). Transposons. In *C.elegans* II, D.L. Riddle, T. Blumenthal, B.J. Meyer, and J.R. Priess, eds. (New York: Cold Spring Harbor Laboratory Press), pp. 97–116.

Poy MN, Eliasson L, Krutzfeldt J, Kuwajima S, Ma X, Macdonald PE, Pfeffer S, Tuschl T, Rajewsky N, Rorsman P, Stoffel M. (2004). A pancreatic islet-specific microRNA regulates insulin secretion. *Nature*. 432(7014), 226-30.

Prachumwat, A., DeVincentis, L., and Palopoli, M.F. (2004). Intron size correlates positively with recombination rate in *Caenorhabditis elegans*. *Genetics* 163, 1585–1590.

Prasad SS, Harris LJ, Baillie DL, Rose AM. (1991). Evolutionarily conserved regions in *Caenorhabditis* transposable elements deduced by sequence comparison. *Genome*. Feb;34(1):6-12.

Reboul, J., Vaglio, P., Rual, J.F., Lamesch, P., Martinez, M., Armstrong, C.M., Li, S., Jacotot, L., Bertin, N., Janky, R., Moore, T., Hudson, J.R. Jr, Hartley, J.L., Brasch, M.A., Vandenhaute, J., Boulton, S., Endress, G.A., Jenna, S., Chevet, E., Papanotiropoulos, V., Tolia, P.P., Ptacek, J., Snyder, M., Huang, R., Chance, M.R., Lee, H., Doucette-Stamm, L., Hill, D.E., and Vidal, M. (2003). *C.elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* 34, 35–41.

Rezsohazy, R., van Luenen, H.G., Durbin, R.M., and Plasterk, R.H. (1997). Tc7, a Tc1-hitch hiking transposon in *Caenorhabditis elegans*. *Nucleic Acids Res.* 25, 4048–54.

Rizzon, C., Martin, E., Marais, G., Duret, L., Segalat, L., and Biemont, C. (2003). Patterns of selection against transposons inferred from the distribution of Tc1, Tc3 and Tc5 insertions in the mut-7 line of the nematode *Caenorhabditis elegans*. *Genetics* 165, 1127–35.

Robertson, H.M. (1998). Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. *Genome Res.* 8, 449–463.

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

Robertson, H.M. (2002). Updating the str and srj (stl) families of chemoreceptors in *Caenorhabditis* nematodes reveals frequent gene movement within and between chromosomes. *Chem. Senses* 26, 151–159.

Robertson, H.M., and Lampe, D.J. (1995). Recent horizontal transfer of a mariner transposable element among and between Diptera and Neuroptera. *Mol. Biol. Evol.* 12, 850–862.

- Rosenzweig, B., Liao, L.W., and Hirsh, D. (1983). Sequence of the *C.elegans* transposable element Tc1. *Nucleic Acids Res.* 11, 4201–9.
- Ruan, K.S., and Emmons, S.W. (1987). Precise and imprecise somatic excision of the transposon Tc1 in the nematode *C.elegans*. *Nucleic Acids Res.* 15, 6875–81.
- Ruvolo, V., Hill, J.E., and Levitt, A. (1992). The Tc2 transposon of *Caenorhabditis elegans* has the structure of a self-regulated element. *DNA Cell Biol.* 11, 111–122.
- Sedensky, M.M., Hudson, S.J., Everson, B., and Morgan, P.G. (1994). Identification of a mariner-like repetitive sequence in *C.elegans*. *Nucleic Acids Res.* 22, 1719–23.
- Sharakhov, I.V., Serazin, A.C., Grushko, O.G., Dana, A., Lobo, N., Hillenmeyer, M.E., Westerman, R., Romero-Severson, J., Costantini, C., Sagnon, N., et al. (2002). Inversions and gene order shuffling in *Anopheles gambiae* and *A. funestus*. *Science* 298, 182–185.
- Sijen, T., and Plasterk, R.H. (2003). Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. *Nature* 426, 310–314.
- Simpson, V.J., Johnson, T.E., and Hammen, R.F. (1986). *Caenorhabditis elegans* DNA does not contain 5-methylcytosine at any time during development. *Nucleic Acids Res.* 14, 6711–19.
- Song, J.J., Smith, S.K., Hannon, G.J., and Joshua-Tor, L. 2004. Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* 305, 1434–37.
- Spieth, J. and Lawson, D. Overview of gene structure (January 18, 2006), *WormBook*, ed. The *C.elegans* Research Community, *WormBook*, doi/10.1895/wormbook.1.65.1, <http://www.wormbook.org>.
- Spieth, J., Brooke, G., Kuersten, S., Lea, K., and Blumenthal, T. (1993). Operons in *C.elegans*: polycistronic mRNA precursors are processed by trans splicing of SL2 to downstream coding regions. *Cell* 73, 521–532.
- Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. (2003). The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* 1, E45.
- Surzycki, S.A., and Belknap, W.R. (2000). Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. *Proc. Natl. Acad. Sci. USA* 97, 245–9.

- Synder, M., and Gerstein, M (2003). Defining genes in the genomics era. *Science* 300, 258–260.
- Tabara, H., Sarkissian, M., Kelly, W.G., Fleenor, J., Grishok, A., Timmons, L., Fire, A., and Mello, C.C. (1999). The *rde-1* gene, RNA interference, and transposon silencing in *C.elegans*. *Cell* 99, 123–132.
- The *C.elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C.elegans*: A platform for investigating biology. *Science* 282, 2012–2018.
- The International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- Tijsterman, M., Ketting, R.F., Okihara, K.L., Sijen, T., and Plasterk, R.H. (2002). RNA helicase MUT-14-dependent gene silencing triggered in *C.elegans* by short antisense RNAs. *Science* 295, 694–7.
- Tops, B.B., Tabara, H., Sijen, T., Simmer, F., Mello, C.C., Plasterk, R.H., and Ketting, R.F. (2005). RDE-2 interacts with MUT-7 to mediate RNA interference in *Caenorhabditis elegans*. *Nucleic Acids Res.* 33, 347–355.
- Tosi, L.R., and Beverley, S.M. (2000). cis and trans factors affecting Mos1 mariner evolution and transposition in vitro, and its potential for functional genomics. *Nucleic Acids Res.* 28, 784–790.
- Troemel, E.R. (1999). Chemosensory signaling in *C.elegans*. *Bioessays* 21, 1011–1020.
- Tu, Z., and Shao, H. (2002). Intra- and inter-specific diversity of Tc3-like transposons in nematodes and insects and implications for their evolution and transposition. *Gene* 282, 133–142.
- Vaglio, P., Lamesch, P., Reboul, J., Rual, J.F., Martinez, M., Hill, D., and Vidal, M. (2003). WormDB: the *Caenorhabditis elegans* ORFeome Database. *Nucleic Acids Res.* 31, 237–240.
- van Luenen, H.G., Colloms, S.D., and Plasterk, R.H. (1993). Mobilization of quiet, endogenous Tc3 transposons of *Caenorhabditis elegans* by forced expression of Tc3 transposase. *EMBO J.* 12, 2513–20.
- van Luenen, H.G., Colloms, S.D., and Plasterk, R.H. (1994). The mechanism of transposition of Tc3 in *C.elegans*. *Cell* 79, 293–301.
- van Pouderooyen, G., Ketting, R.F., Perrakis, A., Plasterk, R.H., and Sixma, T.K. (1997). Crystal structure of the specific DNA-binding domain of Tc3 transposase of *C.elegans* in complex with transposon DNA. *EMBO J.* 16, 6044–54.

- Vastenhouw, N.L., and Plasterk, R.H. (2004). RNAi protects the *Caenorhabditis elegans* germline against transposition. *Trends Genet* 20, 314–319.
- Vastenhouw, N.L., Fischer, S.E., Robert, V.J., Thijssen, K.L., Fraser, A.G., Kamath, R.S., Ahringer, J., and Plasterk, R.H. (2003). A genome-wide screen identifies 27 genes involved in transposon silencing in *C.elegans*. *Curr. Biol.* 13, 1311–16.
- Verdel, A., Jia, S., Gerber, S., Sugiyama, T., Gygi, S., Grewal, S.I., and Moazed, D. (2004). RNAi-mediated targeting of heterochromatin by the RITS complex. *Science* 303, 672–6.
- Volpe, T.A., Kidner, C., Hall, I.M., Teng, G., Grewal, S.I., and Martienssen, R.A. 2002. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* 297, 1833–37.
- Vos, J.C., and Plasterk, R.H. (1994). Tc1 transposase of *Caenorhabditis elegans* is an endonuclease with a bipartite DNA binding domain. *EMBO J.* 13, 6125–32.
- Vos, J.C., De Baere, I., and Plasterk, R.H. (1996). Transposase is the only nematode protein required for in vitro transposition of Tc1. *Genes Dev.* 10, 755–761.
- Vos, J.C., van Luenen, H.G., and Plasterk, R.H. (1993). Characterization of the *Caenorhabditis elegans* Tc1 transposase in vivo and in vitro. *Genes Dev.* 7, 1244–53.
- Ward, S., Burke, D.J., Sulston, J.E., Coulson, A.R., Albertson, D.G., Ammons, D., Klass, M., and Hogan, E. (1988). Genomic organization of major sperm protein genes and pseudogenes in the nematode *Caenorhabditis elegans*. *J Mol Biol.* 199, 1–13.
- Waterhouse P.M., Graham, M.W., Wang, M.B. (1998). Virus resistance and gene silencing in plants can be induced by simultaneous expression of sense and antisense RNA. *Proc Natl Acad Sci U S A.* 95(23), 13959-64.
- Watkins, S., van Pouderooyen, G., and Sixma, T.K. (2004). Structural analysis of the bipartite DNA-binding domain of Tc3 transposase bound to transposon DNA. *Nucleic Acids Res.* 32, 4306–12.
- Whitton, C., Daub, J., Quail, M., Hall, N., Foster, J., Ware, J., Ganatra, M., Slatko, B., Barrell, B., and Blaxter, M. (2004). A genome sequence survey of the filarial nematode *Brugia malayi*: repeats, gene discovery, and comparative genomics. *Mol. Biochem. Parasitol.* 137, 215–227.
- Youngman, S., van Luenen, H.G., and Plasterk, R.H. (1996). Rte-1, a retrotransposon-like element in *Caenorhabditis elegans*. *FEBS Lett.* 380, 1–7.

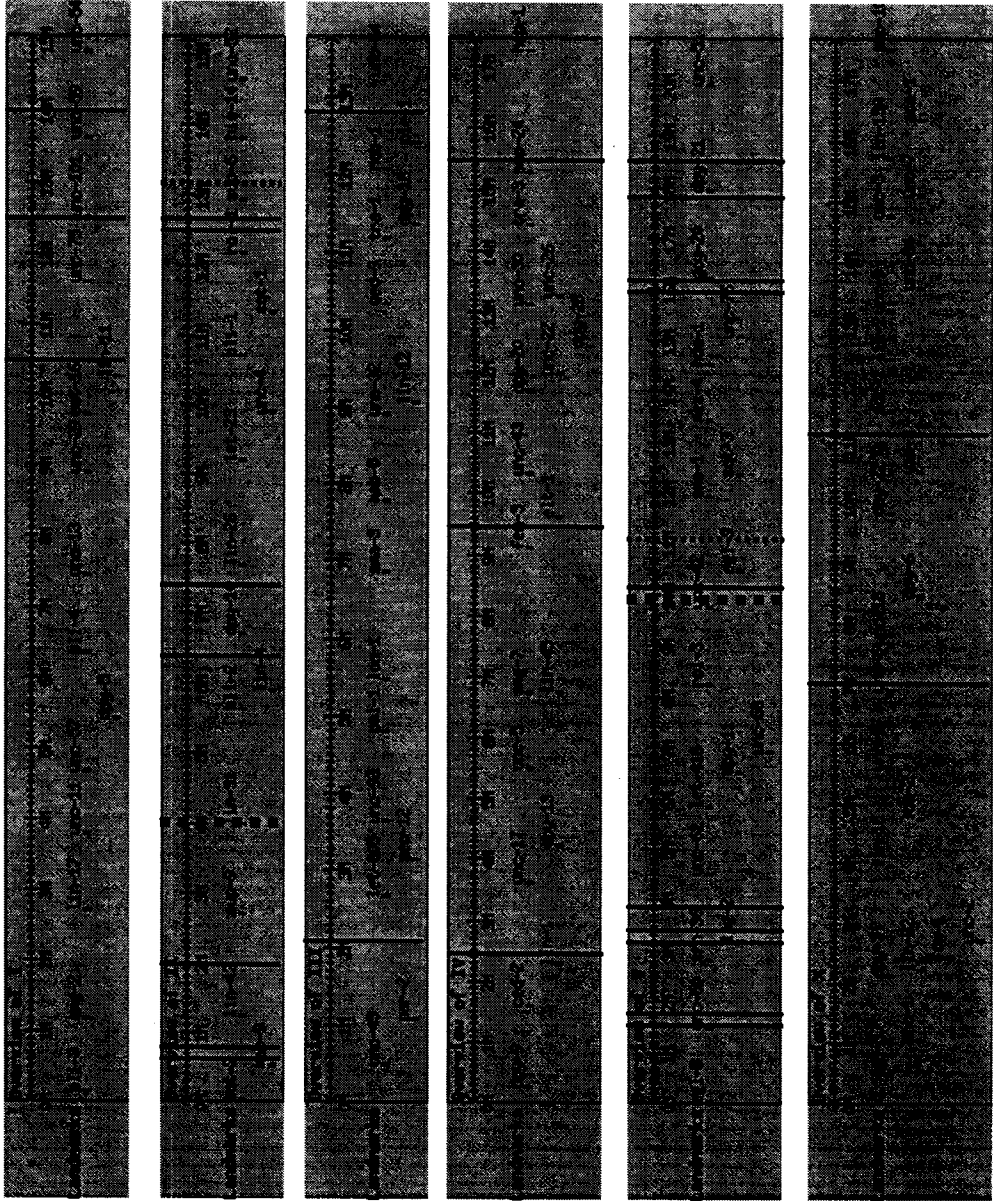
- Yuan, J.Y., Finney, M., Tsung, N., and Horvitz, H.R. (1991). Tc4, a *Caenorhabditis elegans* transposable element with an unusual fold-back structure. *Proc. Natl. Acad. Sci. USA* 88, 3334–38.
- Zagobelny, M., Jeffares, D.C., and Arctander, P. (2004). Differences in non-LTR retrotransposons within *C.elegans* and *C.briggsae* genomes. *Gene* 330, 61–66.
- Zagulski, M., Nowak, J.K., Le Mouel, A., Nowacki, M., Migdalski, A., Gromadka, R., Noel, B., Blanc, I., Dessen, P., Wincker, P., Keller, A.M., Cohen, J., Meyer, E., and Sperling, L. (2004). High coding density on the largest *Paramecium tetraurelia* somatic chromosome. *Curr. Biol.* 14, 1397–1404.
- Zayed, H., Izsvak, Z., Khare, D., Heinemann, U., and Ivics, Z. (2003). The DNA-bending protein HMGB1 is a cellular cofactor of Sleeping Beauty transposition. *Nucleic Acids Res.* 31, 2313–22.
- Zhang H, Emmons S.W. (2000). A *C.elegans* mediator protein confers regulatory selectivity on lineage-specific expression of a transcription factor gene. *Genes Dev.* 14(17), 2161-72.
- Zorio, D.A., Cheng, N.N., Blumenthal, T., and Spieth, J. (1994). Operons as a common form of chromosomal organization in *C.elegans*. *Nature* 372, 270–272.

APPENDICES

APPENDIX A: LINKAGE GROUP MAPS

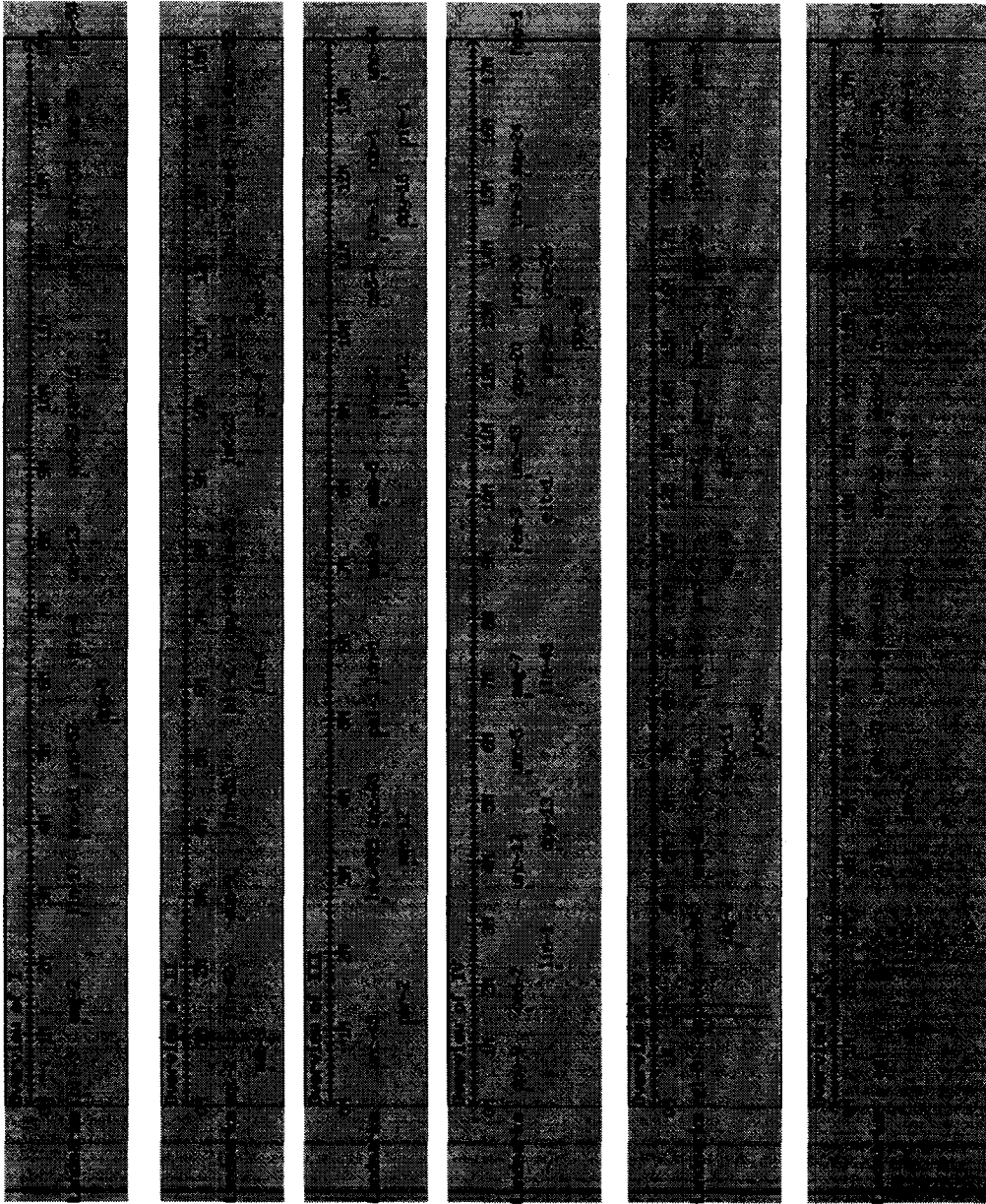
Following are maps that approximate all the full and partial fragment *C.elegans* transposable elements on linkage group maps utilized by permission from WormBase. Each element family (Tc1-10, except 8) is a different color. Solid lines indicate full elements and dashed lines indicate fragments. Additionally, the approximate width of the line indicated numbers of elements in a particular region, i.e. there are up to three different line thicknesses displayed, representing one, 2 or 3 fragments in that region respectively. Following the identifier at the top of each page (Tc1 for example) there is a number preceding "F" to indicate the number of full elements of this type, and a number preceding "P" to indicate the number of partial fragment matches for this element.

TC1 (27 F, 6P)

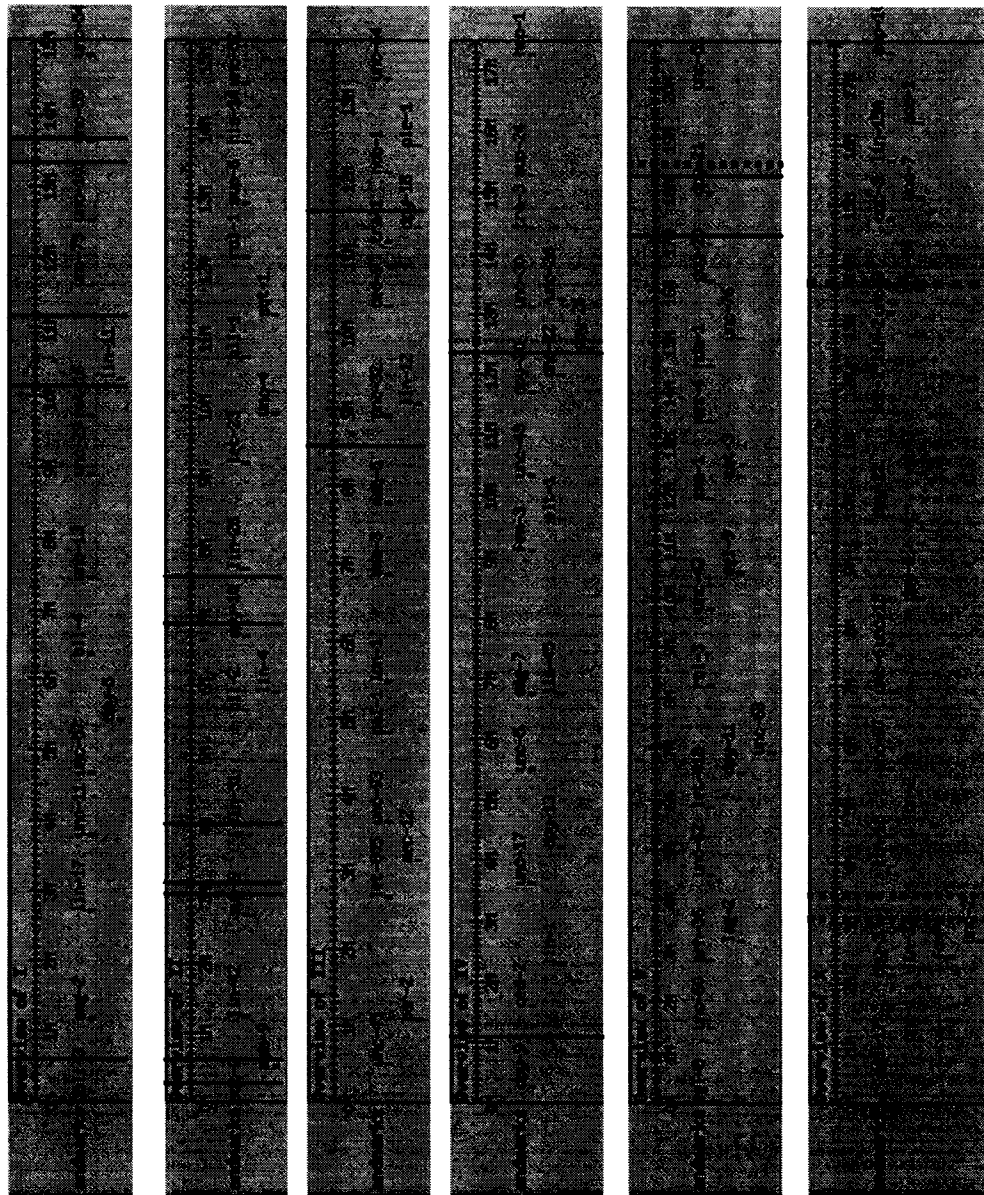


FP 2P

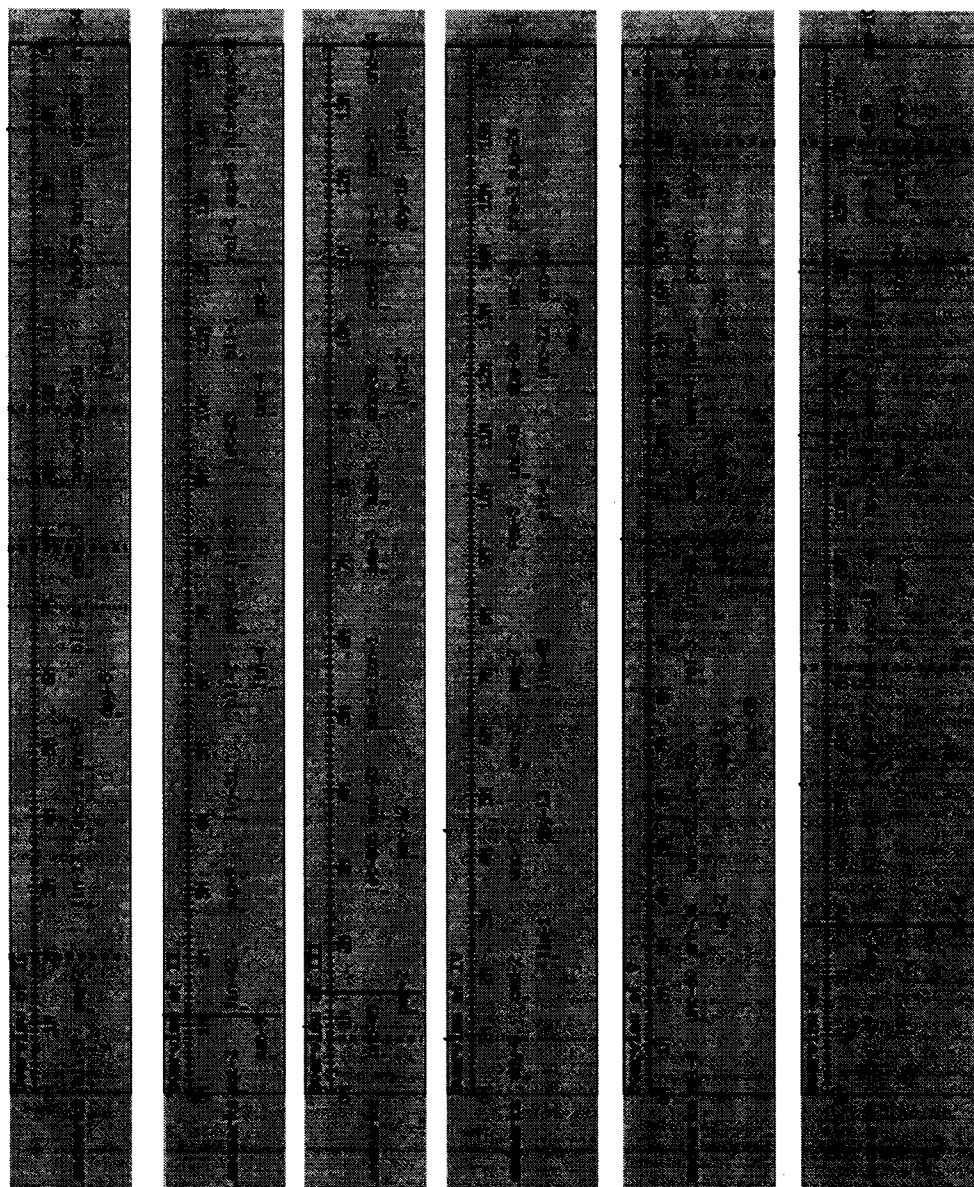
TC2 (4F, 0P)

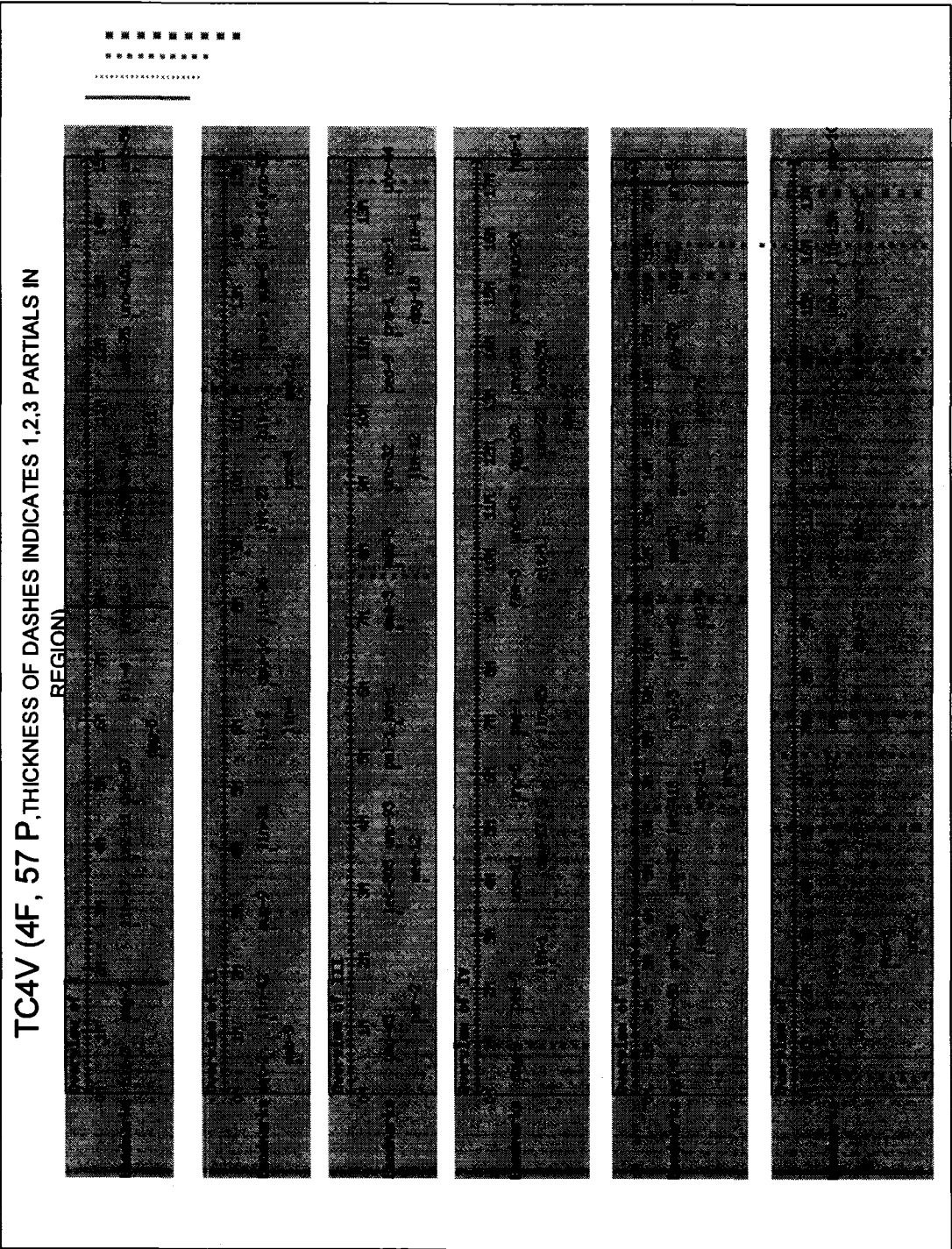


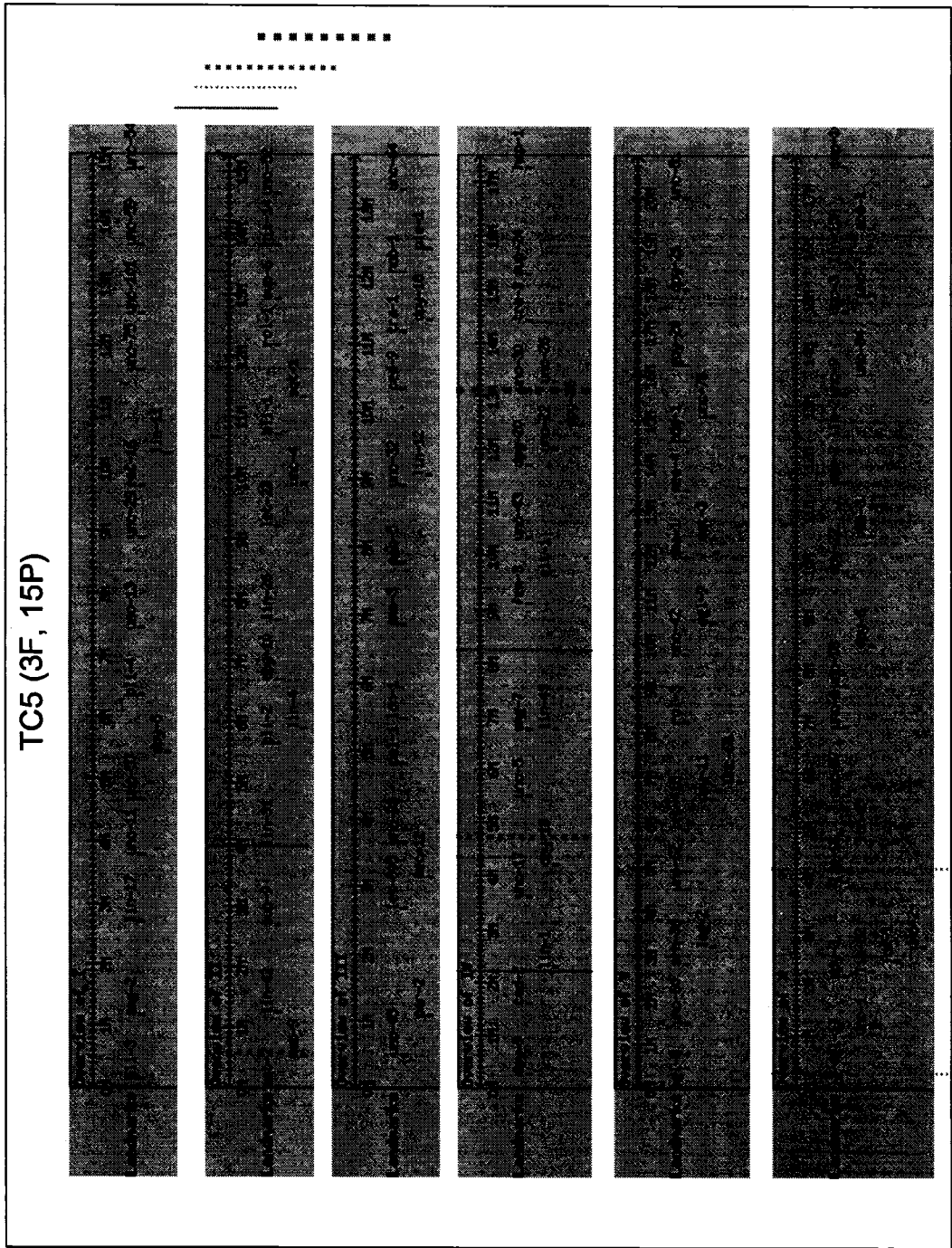
TC3 (18F, 11P) (PARTIALS RIGHT NEXT TO ONE ANOTHER)



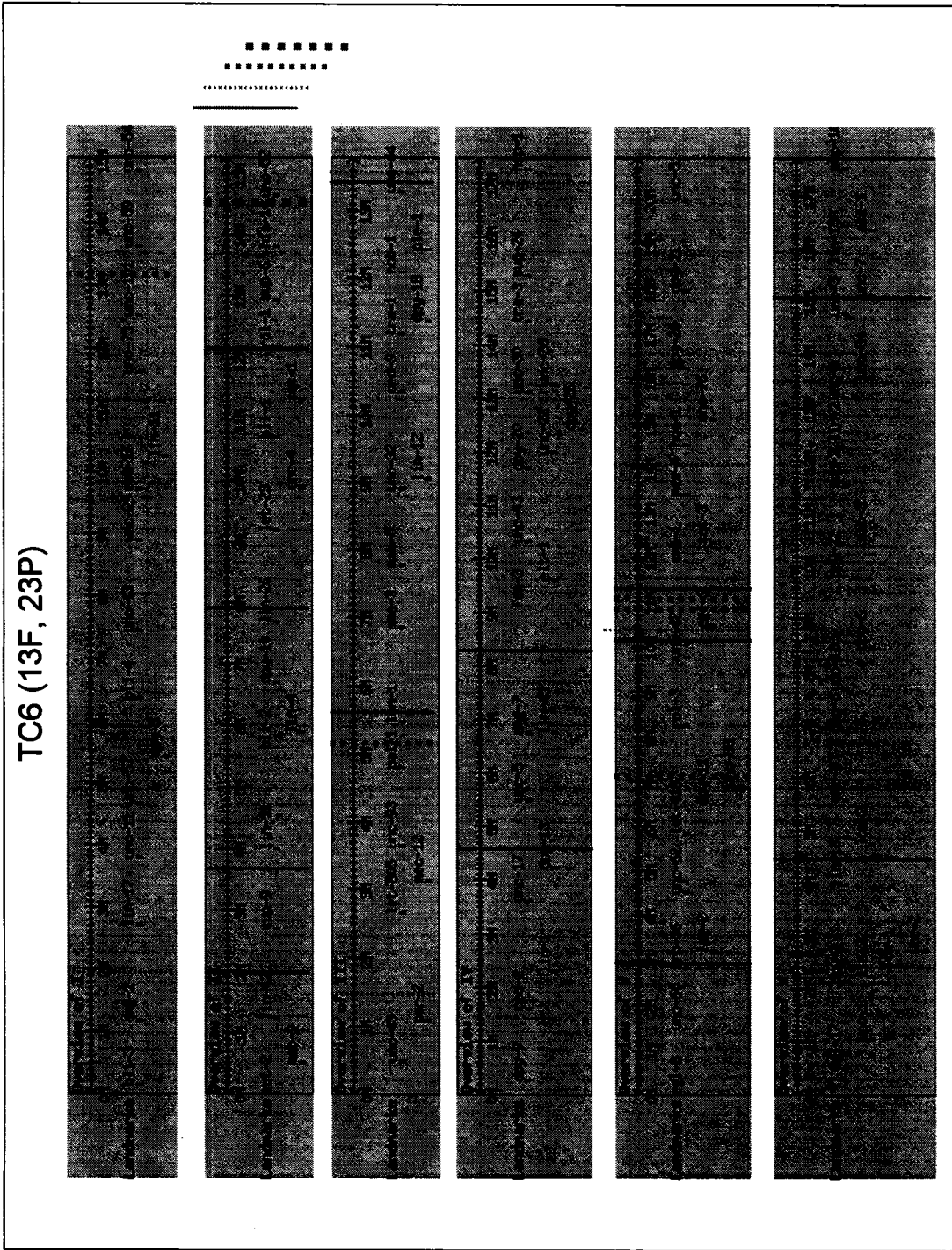
TC4 (3F, 29P, DOUBLE WIDTH INDICATES PAIR OF PARTIALS)



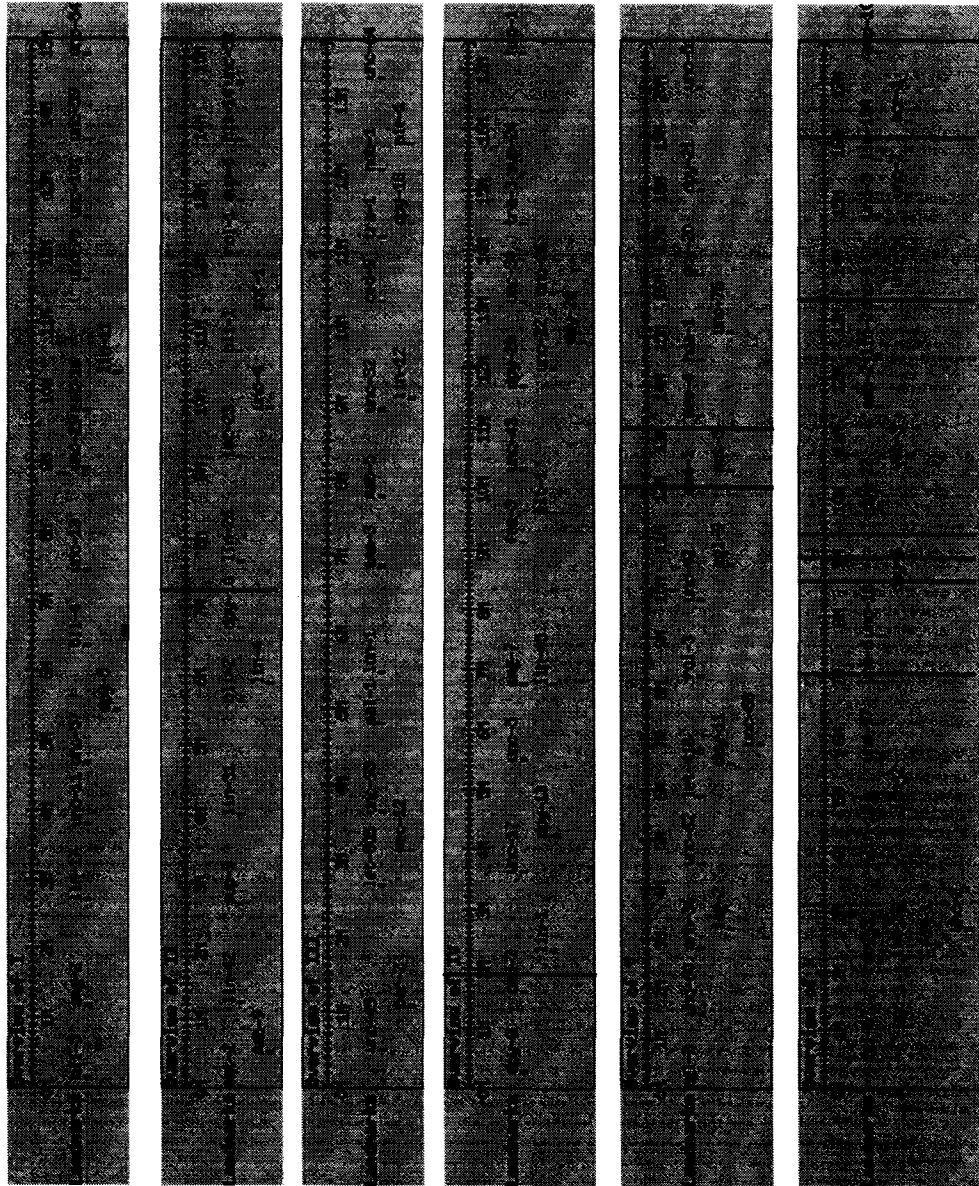




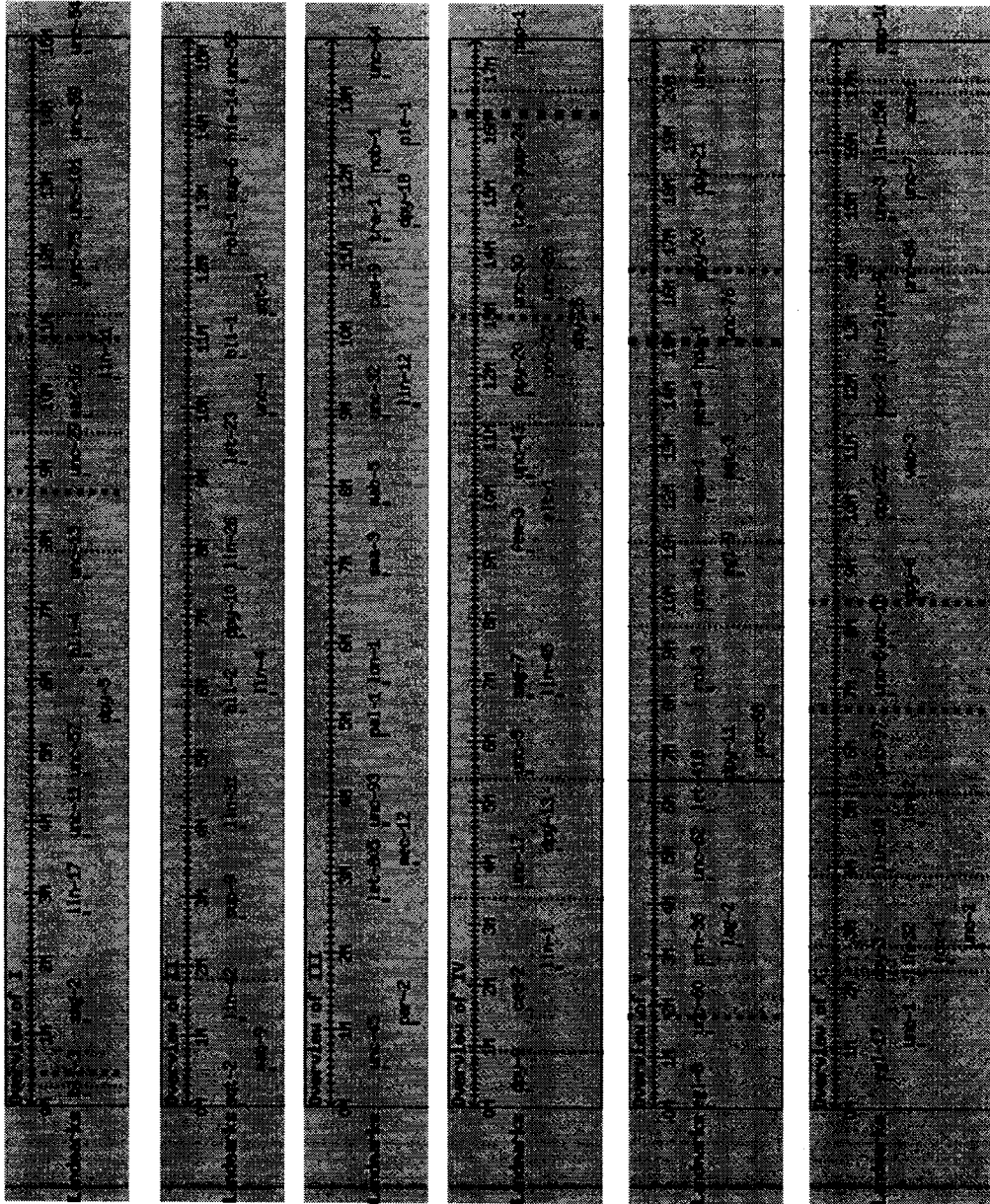
TC5 (3F, 15P)



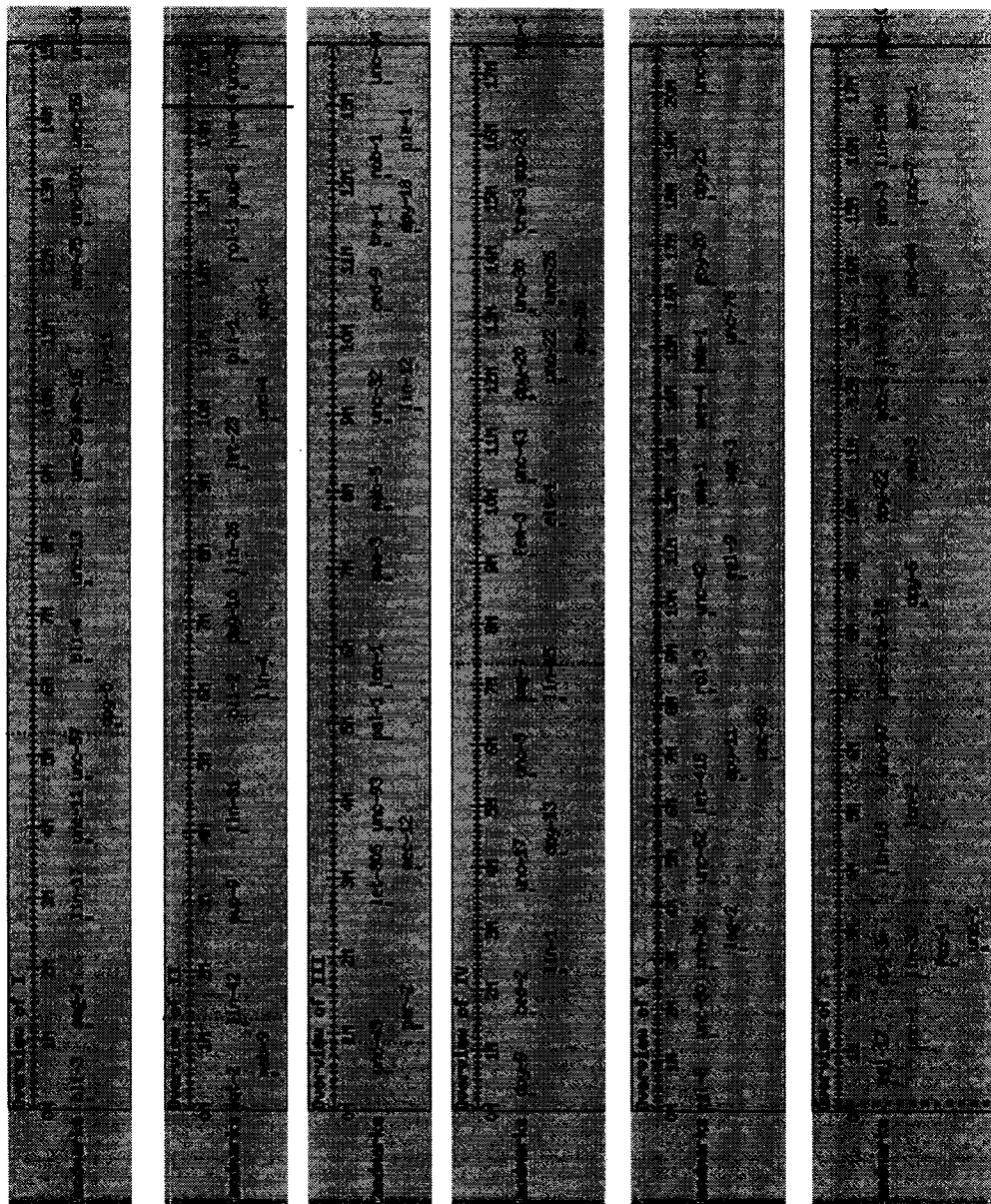
TC7 (10F, 1P)



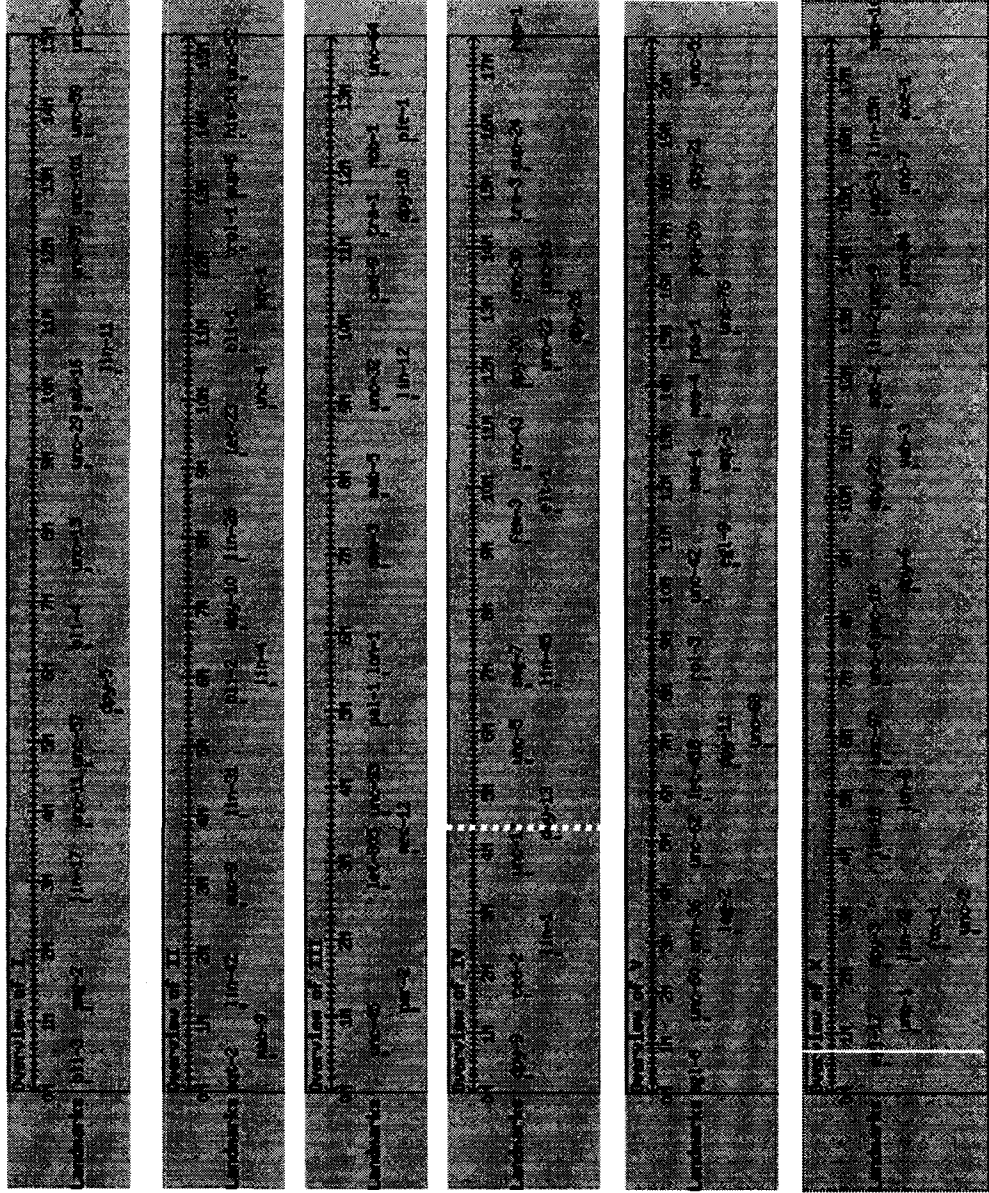
TC9 (1F, 44P)



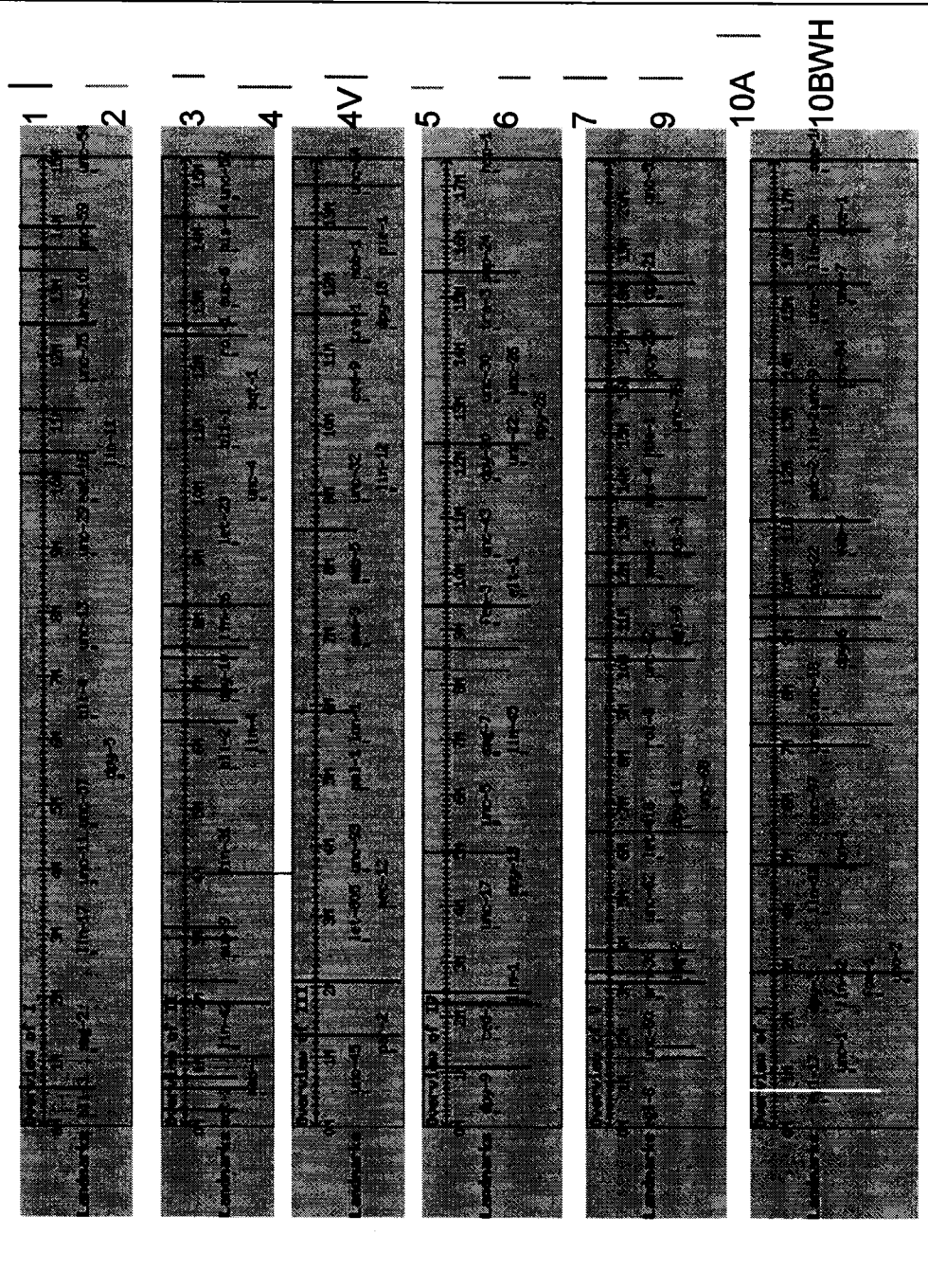
TC10A (1F,7P)



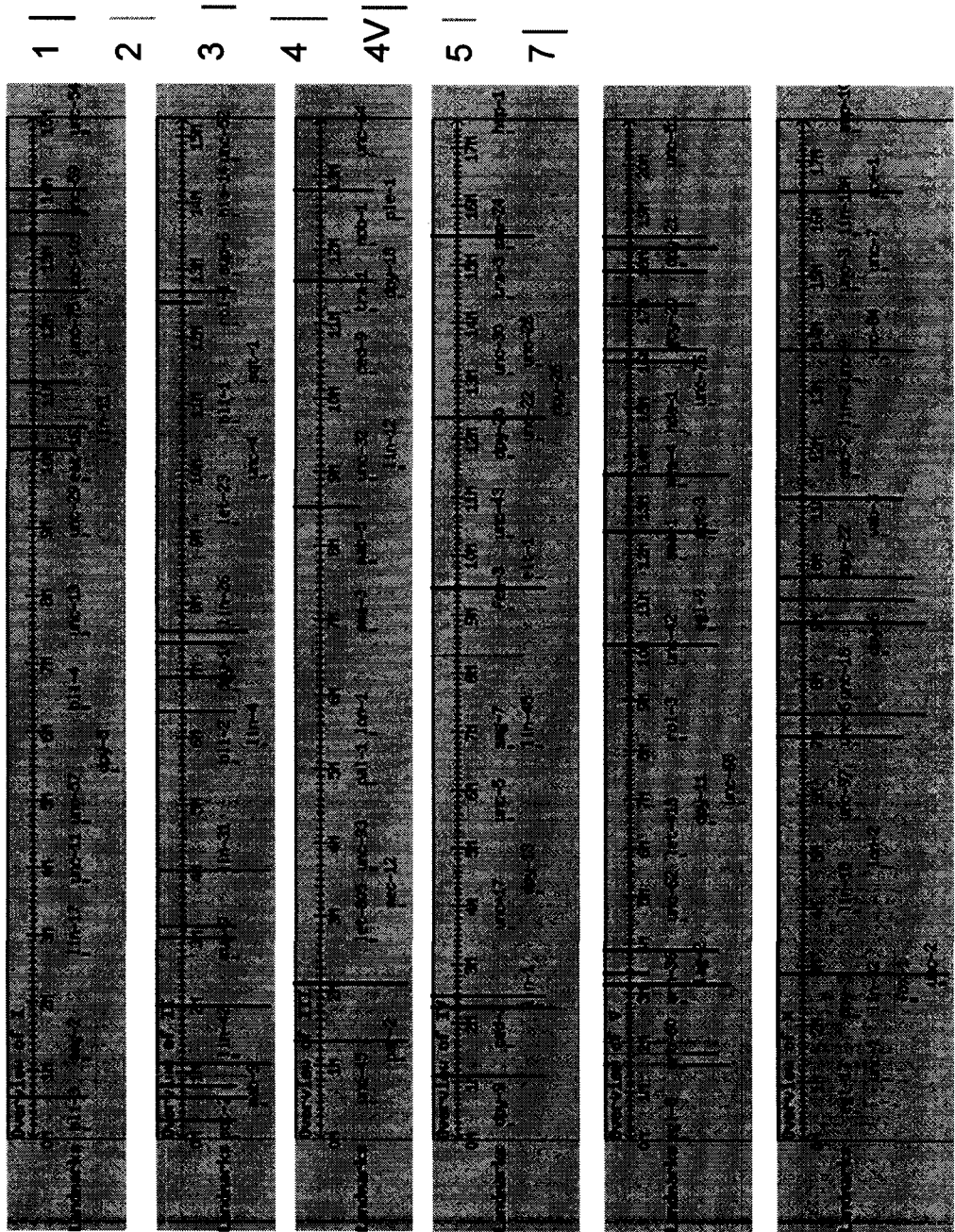
TC10B (1F,2P)



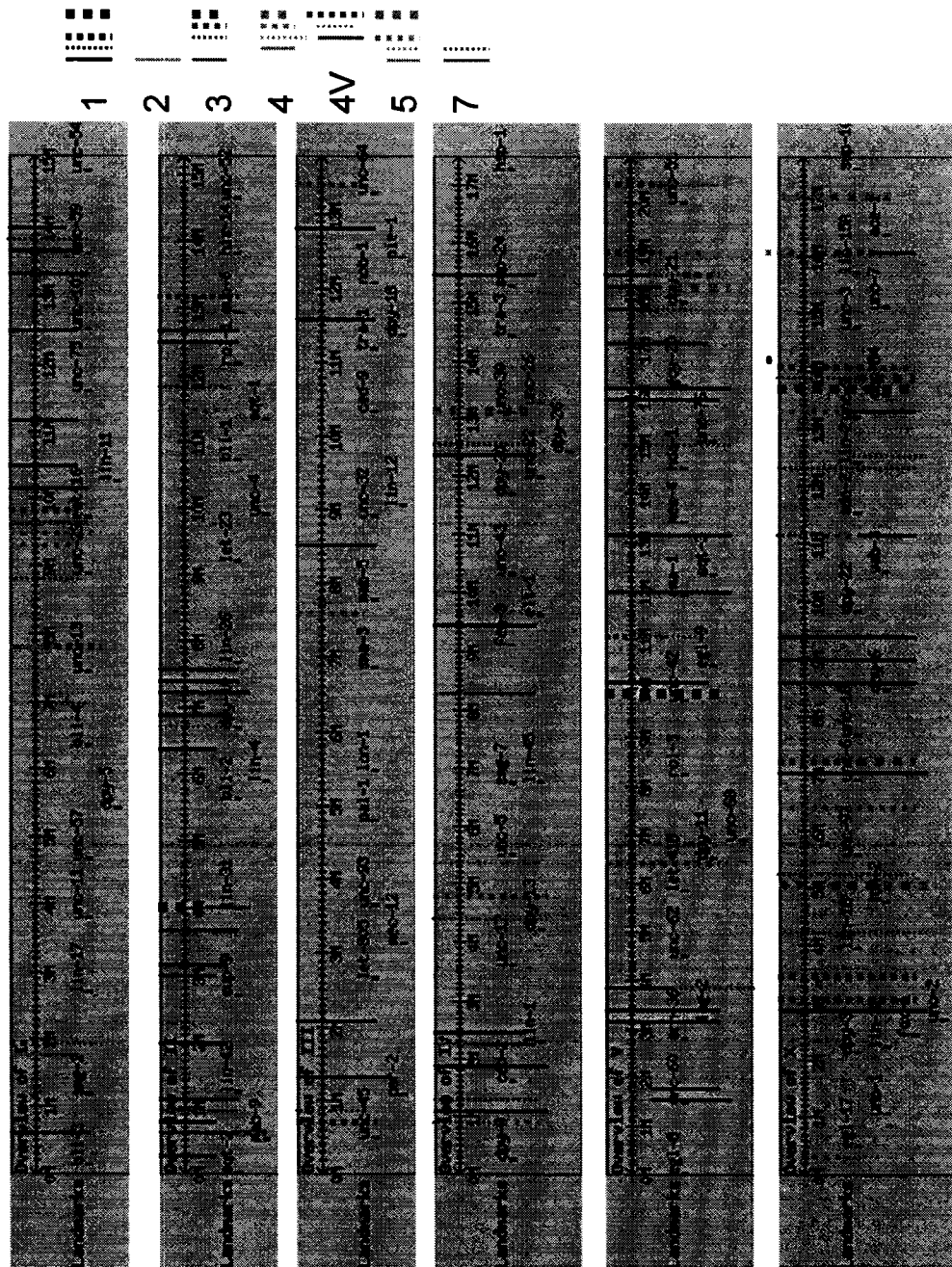
ALL FULL



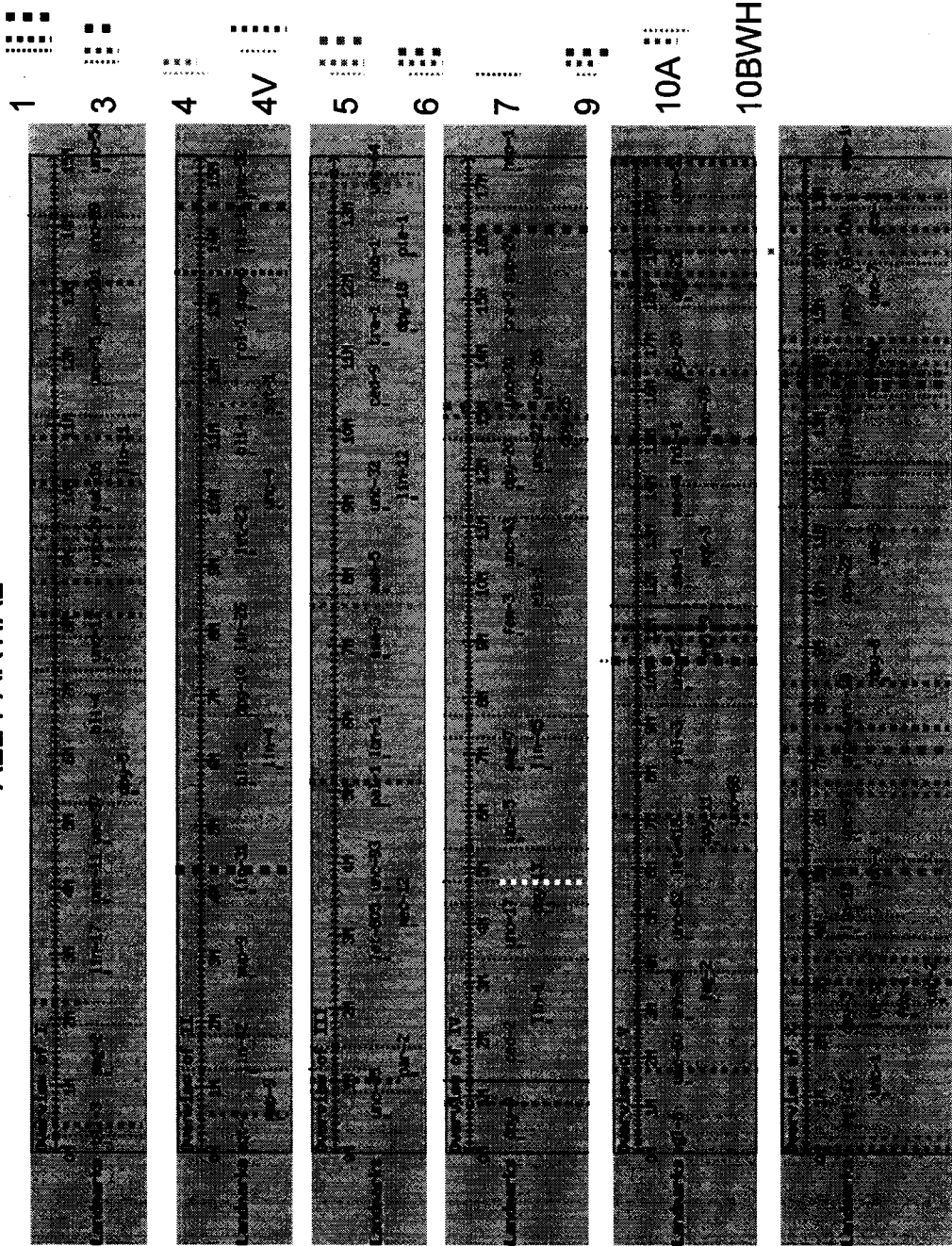
ALL FULL ACTIVE (1,2,3,4,4V,5,7)

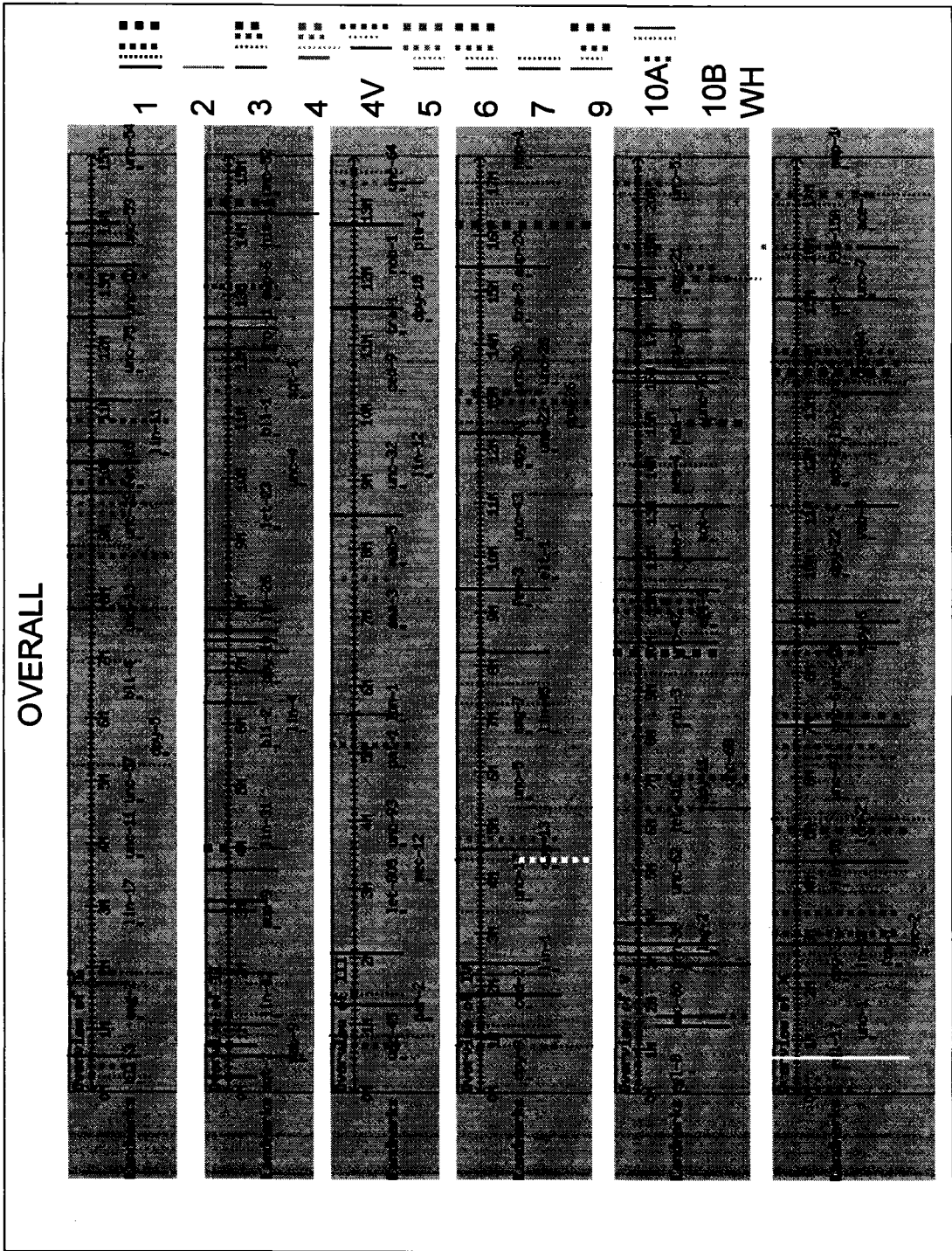


ALL ACTIVE ELEMENTS (1,2,3,4,4V,5,7)



ALL PARTIAL

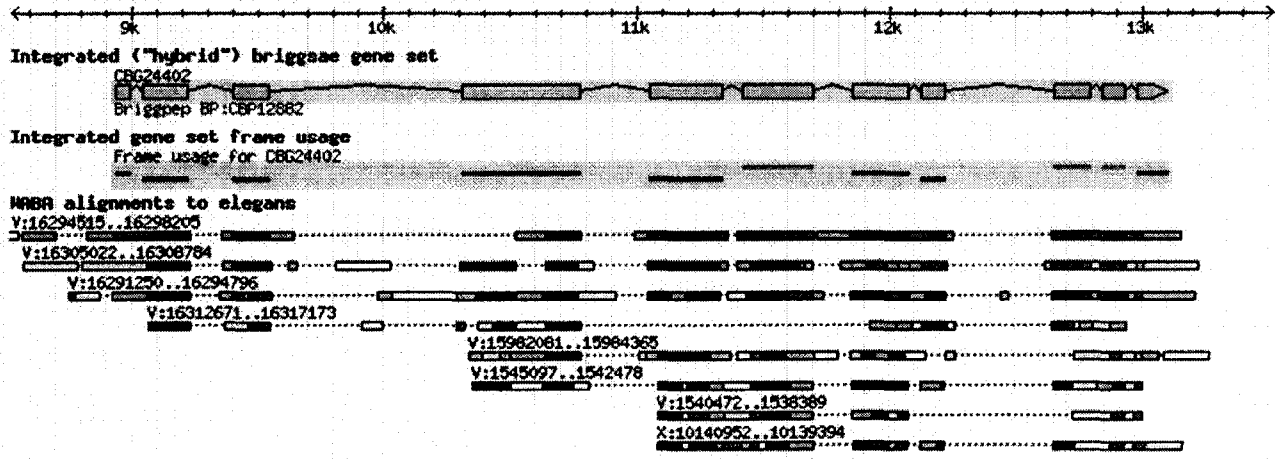
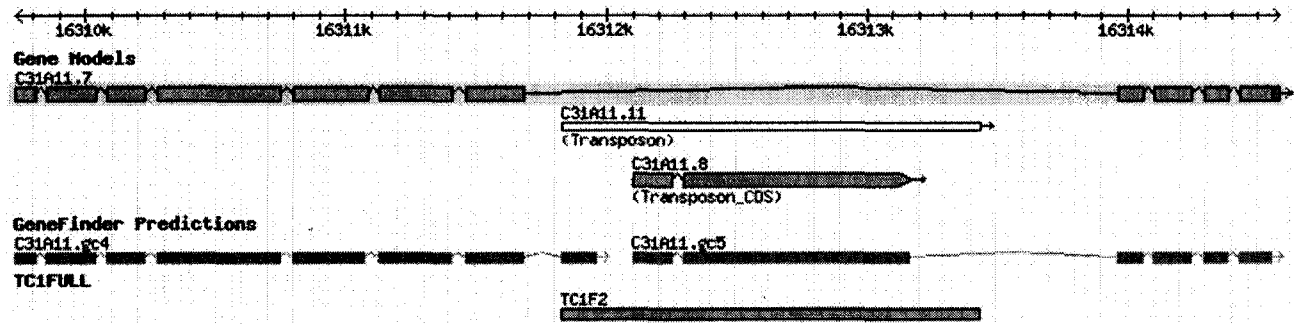




APPENDIX B: INTRON STUDIES

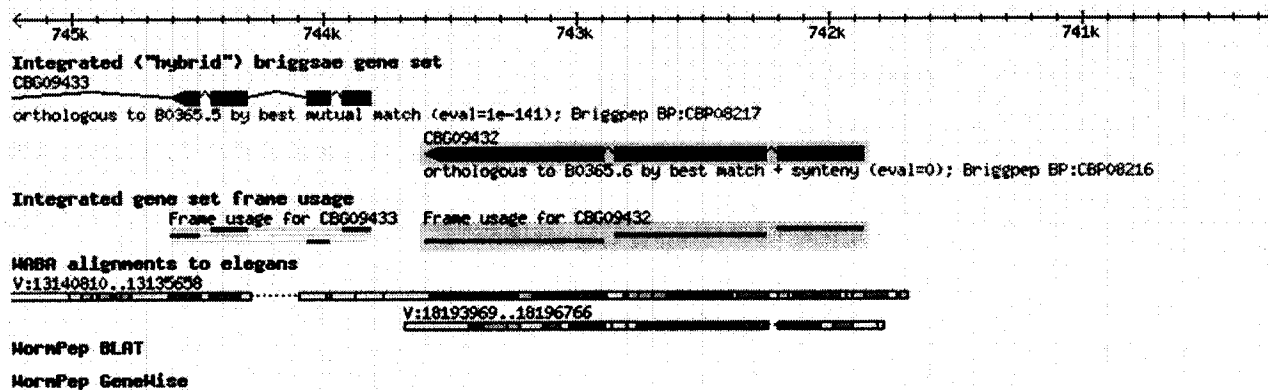
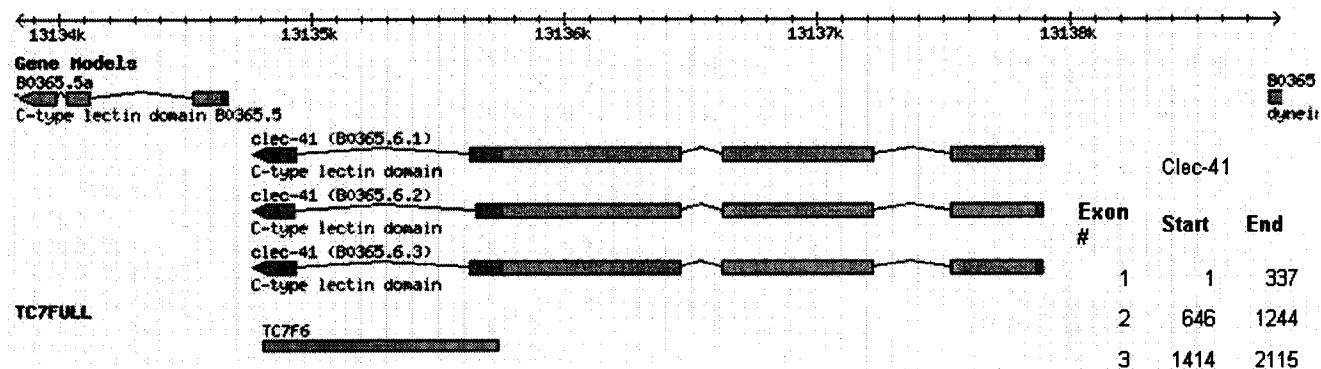
C.elegans Intron Study
(E Intron Study)

Element matches 2098-3709,(between exon 7&8 (1734-1952)(4235-4329)



WormPep BLAT

element matches 2158-3080 – element on other strand, after exon 3(final coding exon) and into intron and non coding exon 4 and into utr, ortholog by wormbase



C elegans transcript

```

1621 TCTGGCCAAGGTTTCAGGATTCAGTGCAAATTTTGGGCGCTCTAAATTTTCCTTGAATCT
1594 TCTGGCCAAGGTTTCAGGATTCAGTGCAAATTTTGGGCGCTCTAA.....
532 -S-G-Q-G-S-G-F-S-A-N-F-W-A-L-*.....

1681 ACAGTGCTGGCCAAAAAGATATCCACTTTCTGTTTTTGGATGATTTTCGATATTTTTTCCA
.....
.....

1741 TTGAGCATAACTCCAAAAGTAGGAAAGCTATCAAAAAGTTTCAACTGGTTCGACCTACAT
.....
.....

1801 TTTACCAGTTGAAAATTTTTTGGTACCTTTCCTAGTTTGAAGTTATGCCATTGGAAAA
.....
.....

1861 AATATCGAAATCGTCAAAAAGTAGGAAAGCTATCTTTTGGCCAGCACTGTAATTATT
.....
.....

1921 TTTCGTGGTGTCTAATTCTTGAATTACTGGTCAATAA

```

123

	490	500	510	520	530	540	
479	AIDTEQCCEI	IEYEDGPLLGSPI	LOIVSG	YPAHAKITQS	SFSMLV	F TDS	SDQG GF YBGene0000
480	AIDTEQCCEI	IVY DGPLLGS	V LOIVSG	YPAESHITKS	SFSMLV	F TD	SDQG GF CBG09432
	AIDTEQCCEI	-YrDGPLLGSq	-LOIVSG	WPAH-K-Y	-StSFSMLV	FtTD	-SDQQtGF consensus
539	S	EFVAL					YBGene0000
540	S	ITAT					CBG09432
	Sg	-F-A-					consensus

C elegans/C briggsae protein align.

C elegans (cdna+(genomic-intron)/ C briggsae genomic and predicted cdna

```

2180 2200 2220 2240 2260 2280
1557 AGGACA GCGAAA TAAAGGCTAAGTTCGAAATTCATGCTTTT AAATTTTC ACATTA W30ene0000
2034 AGGACA GCGAAA TAAAGGCTTCGTCGAAATTCATGCTTTT AAATTTTC ACATTA W30ene00_1
1525 AGGACA TCGAAA TAAAGGCTAAGTTCGAAATTCATGCTTTT AAATTTTC ACATTA CB009432_1
1533 AGGACA TCGAAA TAAAGGCTAAGTTCGAAATTCATGCTTTT AAATTTTC ACATTA CB009432_3
AGGACAC-G-AAAActTAT-AGTC--G-TG-AATTCAAATGCTTGT-G-TTC-OgAGAGA consensus

```

```

2130 2150 2170 2190 2210 2230
1517 TTAAGTTCGCAAGGTTTAAAGATTCAGTGAATTTTCGCTTAAATTTTCATTTTGA W30ene0000
2094 TTAAGTTCGCAAGGTTTAAAGATTCAGTGAATTTTCGCTTAAATTTTCATTTTGA W30ene00_1
1585 TTAAGTTCGCAAGGTTTAAAGATTCAGTGAATTTTCGCTTAAATTTTCATTTTGA CB009432_1
1593 TTAAGTTCGCAAGGTTTAAAGATTCAGTGAATTTTCGCTTAAATTTTCATTTTGA CB009432_3
T-G-TGTCG-CAAGC--GAGGATTCAGTC-A-TTC--GC-t-tgatttcctttga consensus

```

```

2170 2190 2210 2230 2250 2270
1577 ATCGAGACTGCTGGCCAAAAGATATCCAGTTCGTCTTTTTCAGATTCGATATTTT W30ene0000
2154 ATCGAGACTGCTGGCCAAAAGATATCCAGTTCGTCTTTTTCAGATTCGATATTTT W30ene00_1
1734 ATCGAGACTGCTGGCCAAAAGATATCCAGTTCGTCTTTTTCAGATTCGATATTTT CB009432_1
1642 atctacagtgctggcctaaaagatattccatttctgttttttgatgatttgatattttt consensus

```

```

2200 2220 2240 2260 2280 2300
1737 TCGTTCAGCATAAAGTGGAAAAGTATGAAAGATATGAAAAGTTTCAARTGCT..... W30ene0000
2214 TCGTTCAGCATAAAGTGGAAAAGTATGAAAGATATGAAAAGTTTCAARTGCT..... W30ene00_1
1734 TCGTTCAGCATAAAGTGGAAAAGTATGAAAGATATGAAAAGTTTCAARTGCT..... CB009432_1
1642 tccattgagcataaactcctaaaactagggaagctatcctaaaagtttcaactggt----- consensus

```

```

2200 2220 2240 2260 2280 2300
1791 ..... W30ene0000
2274 TAACTGATGATGAGGCGTATGTATTTTACATGATTAAATTATTCAGGAAAAGATGCA W30ene00_1
1734 TAACTGATGATGAGGCGTATGTATTTTACATGATTAAATTATTCAGGAAAAGATGCA CB009432_1
1642 TAACTGATGATGAGGCGTATGTATTTTACATGATTAAATTATTCAGGAAAAGATGCA CB009432_3
consensus

```

```

2200 2220 2240 2260 2280 2300
1791 ..... W30ene0000
2334 AATBAYAAAAGCGCCAGATGCTGCTGAGTGGCTTTTATGATGATTAATAAAAGACTG W30ene00_1
1734 AATBAYAAAAGCGCCAGATGCTGCTGAGTGGCTTTTATGATGATTAATAAAAGACTG CB009432_1
1642 AATBAYAAAAGCGCCAGATGCTGCTGAGTGGCTTTTATGATGATTAATAAAAGACTG CB009432_3
consensus

```

```

2200 2220 2240 2260 2280 2300
1791 ..... W30ene0000
2384 TAACTGAGAAAACATATTTTAAATGAAAGTTTTAAAGATTAAGAGATGTTTATATAA W30ene00_1
1734 TAACTGAGAAAACATATTTTAAATGAAAGTTTTAAAGATTAAGAGATGTTTATATAA CB009432_1
1642 TAACTGAGAAAACATATTTTAAATGAAAGTTTTAAAGATTAAGAGATGTTTATATAA CB009432_3
consensus

```

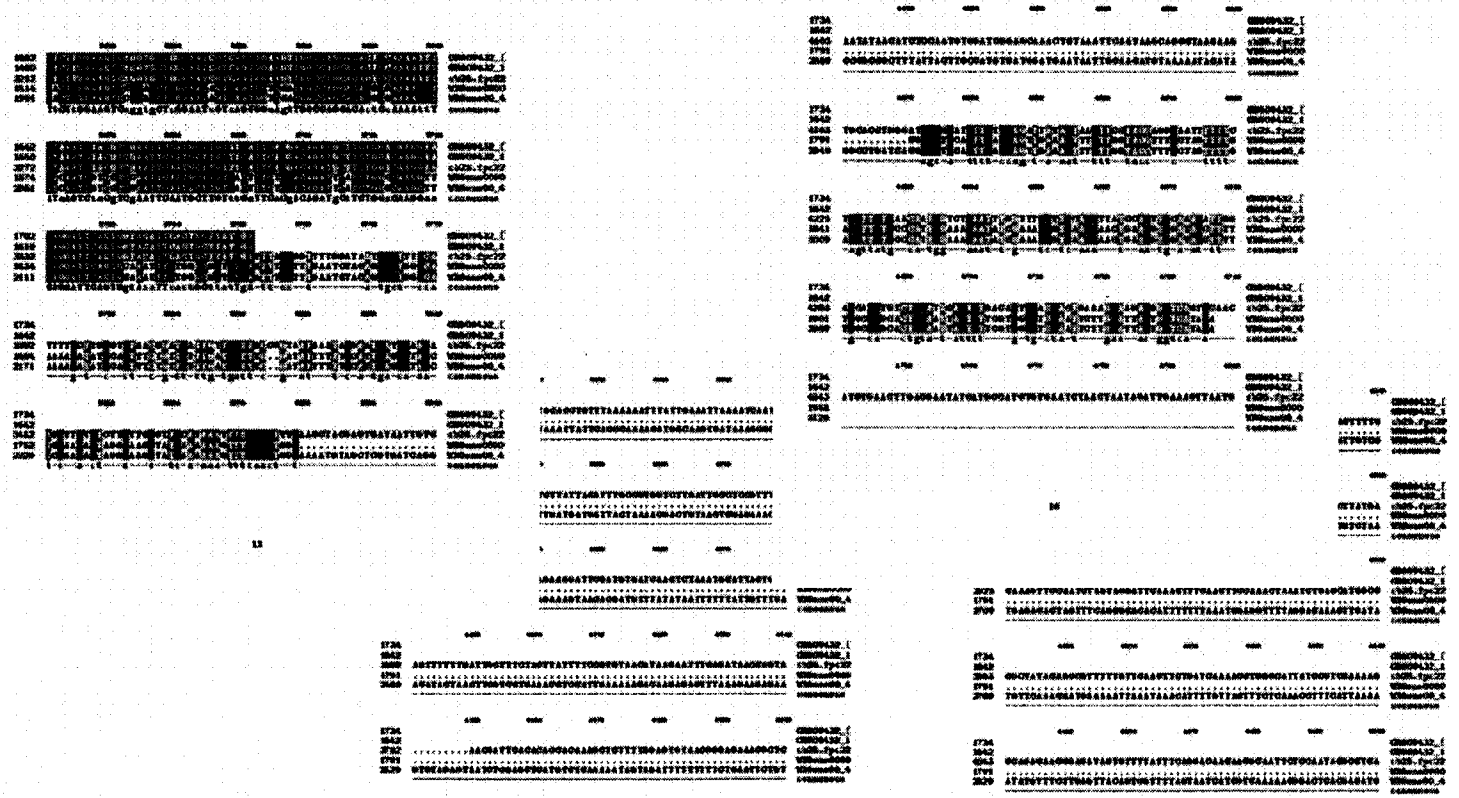
```

2470 2490 2510 2530 2550 2570
1791 ..... W30ene0000
2644 TTTTATATTTTAAAGATGATAAATTTTCTGCTGAAAGCTTCAATTAAGAAAAGAGAA W30ene00_1
1734 TTTTATATTTTAAAGATGATAAATTTTCTGCTGAAAGCTTCAATTAAGAAAAGAGAA CB009432_1
1642 TTTTATATTTTAAAGATGATAAATTTTCTGCTGAAAGCTTCAATTAAGAAAAGAGAA CB009432_3
consensus

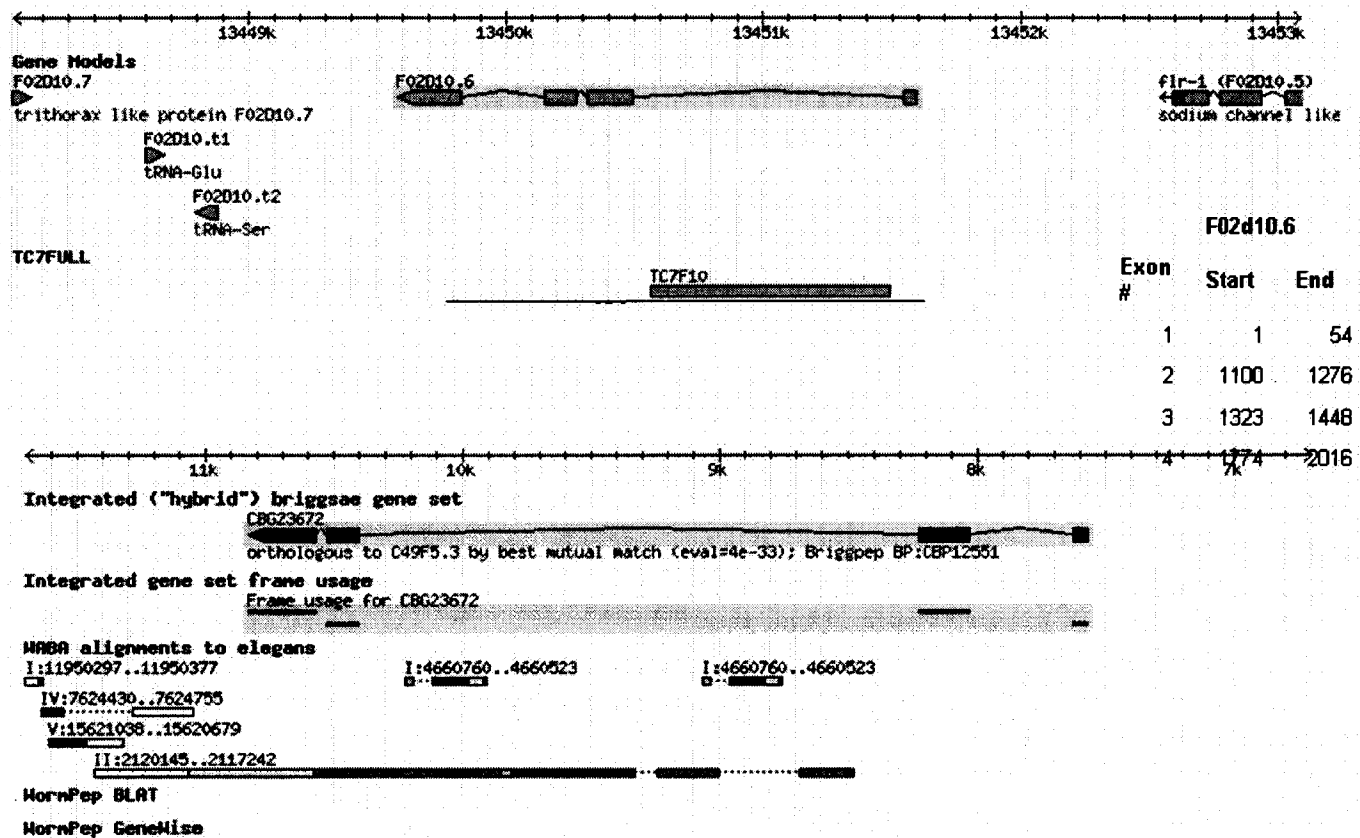
```

alignment with added cb flanking sequence – briqqsae downstream sequence aligns to last exon in elegans gene

125



element matches 104-1025, —on other strand, ortholog by best blast p match



C elegans transcript

```

1 ATGGCAAGTTTCATTGACAAGCTTCCAGAAATCAGCAGGGAAAGCAAGATGGAGATCACC
1 ATGGCAAGTTTCATTGACAAGCTTCCAGAAATCAGCAGGGAAAGCAAGATGGAGATCACC
1 -M--A--S--F--I--D--K--L--P--E--I--S--R--E--S--K--N--E--I--T-

61 CTTCTCTTCTTCTCCTCACTGTCGAAATGGTATATCGTCTGGATTAGCTACTTGTTCATTAAT
61 CTTCTCTTCTTCTCCTCACTGTCGAAATGGTATATCGTCTGGATTAGCTACTTGTTCATTAAT
21 -L--L--F--F--L--T--V--E--M--V--I--V--W--I--S--Y--L--F--I--N-

```

```

1 MASFDKLPEISRESK...VVIS...PVE...FIDN...FV VBGene0000
1 MRSIDGYKPEKRDMEV...LT...TITG...TE...SEWV...ELF...L... CB023672
M-FID-e--S--ME-LL-LTiEiiv---Tly-Ik-Eqsmavvtg-y-Mi-n consensus

```

C elegans/C briggsae protein align.

```

54 ALNLS FGI...FAITRAGPR...GTY...LY...VPPF...SL...F...IT...IT... VBGene0000
60 AAF...VDP...V...RAGPR...L...L...Y...F...Y...V...V...G...L...L...L... CB023672
LiVa-F--i--s--ti--FA-TRAGPRILHls--VVI--s-G-NI-s-fv--fIL-LF-Y-V- consensus

114 SYK...D...T...V...S...L...Q...G...S...R...G...P...H...S...P...I...P...F...I...S...V...A...C...I...T... VBGene0000
120 S...E...R...T...I...G...K...E...T...T...A...T...K...E...N...D...E...N...T...A...V...E...E...D...I...E...S...E...D...V...A...I...Q...N...S...I...Q...S...E... CB023672
q1s-QK-YF-WI-v---Kv---as-dva-r-qf---sE-ssedval--ki-F-E consensus

168 ILL...R...G...Y...A...L...E...D...E...S...P...V...Y...P...D...E...N...S...E...L...I... VBGene0000
180 DS...R...E...D...E...T...E...D...G...E...R...Q...D...A...R...S...E...N...Y...F...E...G...L...E... CB023672
--sDedyktDgy-t-E-ilfd--sas-ki-d--II consen

```

weak protein similarity - not ortholog

```

75 GA...G...A...A...A...C...A...T...G...T...A...G...A...A...G...A...G...G...A...A...A...G...C...A...A...G...A...T...G...G...A...G...A...T...C...A...C... CB023672
481 GA...G...A...A...A...C...A...T...G...T...A...G...A...A...G...A...G...G...A...A...A...G...C...A...A...G...A...T...G...G...A...G...A...T...C...A...C... VBGene0000
1 A...G...A...A...A...C...A...T...G...T...A...G...A...A...G...A...G...G...A...A...A...G...C...A...A...G...A...T...G...G...A...G...A...T...C...A...C... VBGene00_3
1 A...G...A...A...A...C...A...T...G...T...A...G...A...A...G...A...G...G...A...A...A...G...C...A...A...G...A...T...G...G...A...G...A...T...C...A...C... consen
cAt-gc-ag--TOATtG-CACAgct-OcaGA-AtrAcgCAGg-AAAgC-aggcaatccAT consensus

```

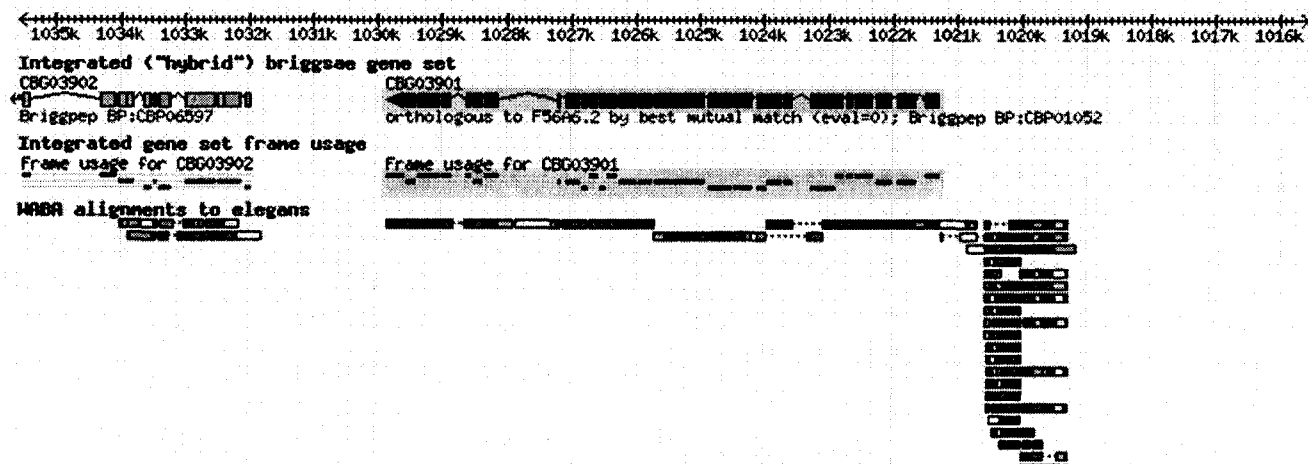
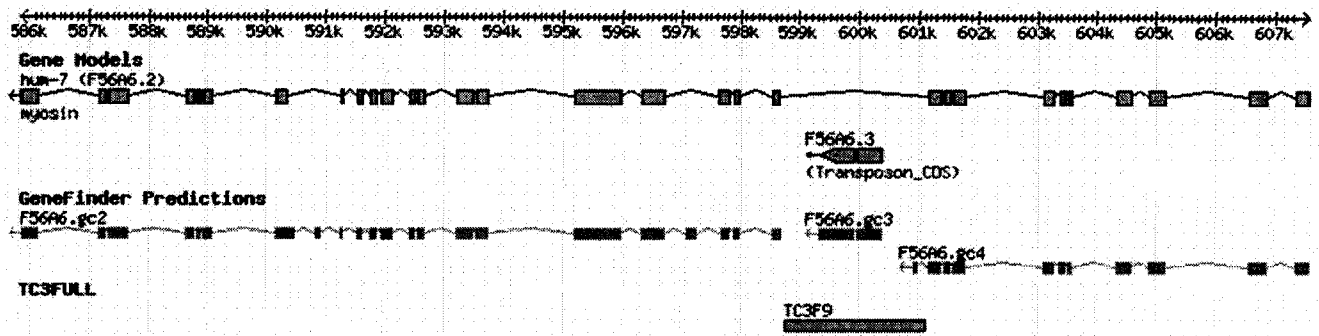
```

135 GGA GTTGTCACTGAGCTATTTCACCTCTTCGGAATCTGGAATCTATTGTGGGCGCTTT CB023672.1
541 GGA GTTGTCACTGAGCTATTTCACCTCTTCGGAATCTGGAATCTATTGTGGGCGCTTT CB023672.1
51 GGA ..... VBGene0000
51 GGA GTA...G...A...T...T...A...T...T...C...A...A...A...A...A...A...G...C...A...A...G...A...T...G...G...A...G...A...T...C...A...C... VBGene00_3
GGAggttgcactgagctatttcacctcttcaggatctgaatctcattgtcgccgctt consensus

```

C elegans (cdna+(genomic-intron)
C briggsae genomic and predicted cdna

element aligns to 6475-8811 (verified by bls2seq in ncbi with intron and tc3 – full alignment), between exon 10&11 (6197-6426)(8913-9011)



C elegans transcript

1801 GAGGTGTGCGAGCTGTTCAGGTAAGTCCAATAAGTCTTNTTGGAGTAAACCATATGGA
 1801 GAGTGTGTCAGCTGTTCAGGTAAGTCCAATAAGTCTTNTTGGAGTAAACCATATGGA
 601 -E-C-V-Q-L-F-Q-V-S-P-I-S-P-F-W-S-K-P-Y-G-

 1861 ATTCAGGCTAGTCGATGAGGAGAGTAATCAATAATGGAACAGATGACTCAATGCTG
 1861 ATTCAGGCTAGTCGATGAGGAGAGTAATCAATAATGGAACAGATGACTCAATGCTG
 621 -I-L-R-L-V-D-E-S-N-I-N-N-G-T-D-D-S-M-L-

C elegans/C briggsae protein align.
 excellent protein alignment, breaks seem to coincide with intron/exon junctions

541 **WIFVATYDHS**
 541 **LRFVFDYV-QHSFEQDIYVAVKELQSLFQIPQFQZVILKRSISYHILEYDRI**
 VS6:ame0000
 CB803901 consensus

 601 **DVNFVSVSE**
 601 **ZVQLQV-p-l-LPTVELLAPEZSISHTDUSHLAKLQFLKHIEYVIZIKLE**
 VS6:ame0000
 CB803901 consensus

 661 **FTVETACKVYQIQWSEYDWRKQVLSNKLSSTISVYVETITQYIAVAVQV**
 660 **YAFVAKLAGEKIQIQWSEYDWRKQVLSNKLSSTISVYVETITQYIAVAVQV**
 VS6:ame0000
 CB803901

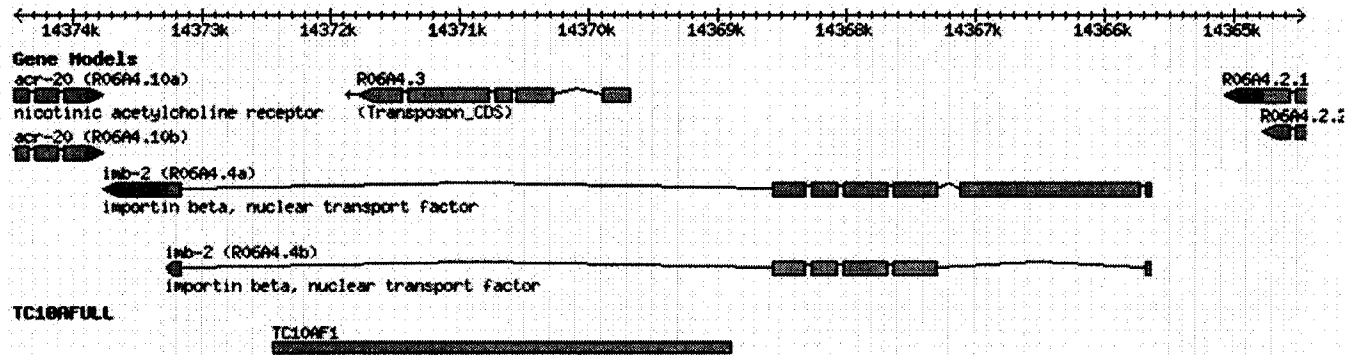
3073 **AARACCA**
 758 **AARACCA**
 473 **AARACCA**
 577 **AARACCA**
 1--A--AGTGTGTS--G-T-It--gagaggttgaagttc--aa-aa-g-ct-
 3099 **CGATGCTG**
 2580 **CGATGCTG**
 523 **CGATGCTG**
 3055 **CGATGCTG**
 VS6:ame0000
 VS6:ame0000
 VS6:ame0000
 VS6:ame00_3 consensus
 3099 **CGATGCTG**
 2580 **CGATGCTG**
 523 **CGATGCTG**
 3055 **CGATGCTG**
 VS6:ame0000
 VS6:ame0000
 VS6:ame00_3 consensus

C elegans (cdna+genomic-intron)/C briggsae genomic and predicted cdna

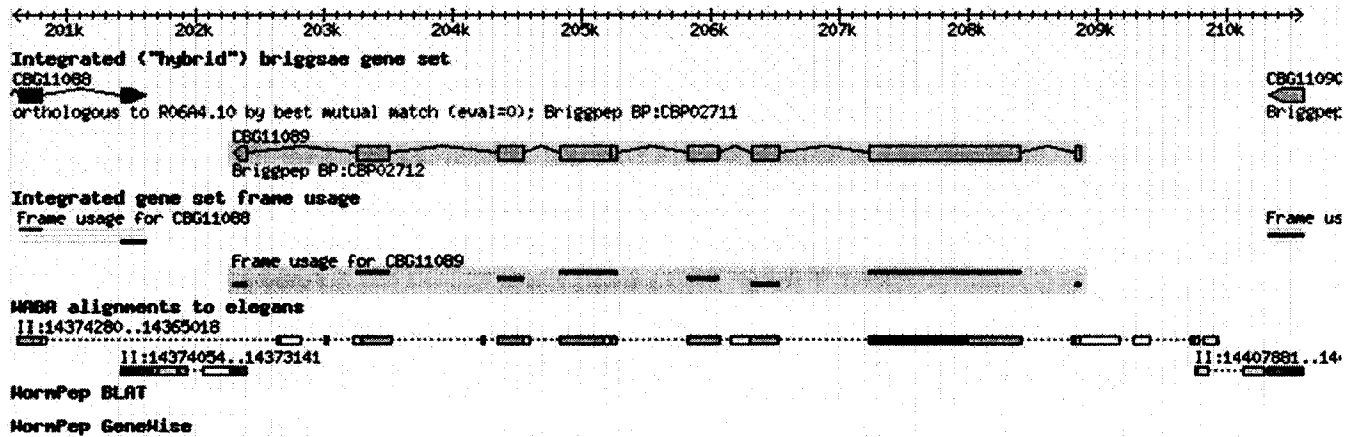
b has exons and introns in e element intron region

3099 **CGATGCTG**
 2580 **CGATGCTG**
 523 **CGATGCTG**
 3055 **CGATGCTG**
 VS6:ame0000
 VS6:ame0000
 VS6:ame00_3 consensus
 3099 **CGATGCTG**
 2580 **CGATGCTG**
 523 **CGATGCTG**
 3055 **CGATGCTG**
 VS6:ame0000
 VS6:ame0000
 VS6:ame00_3 consensus

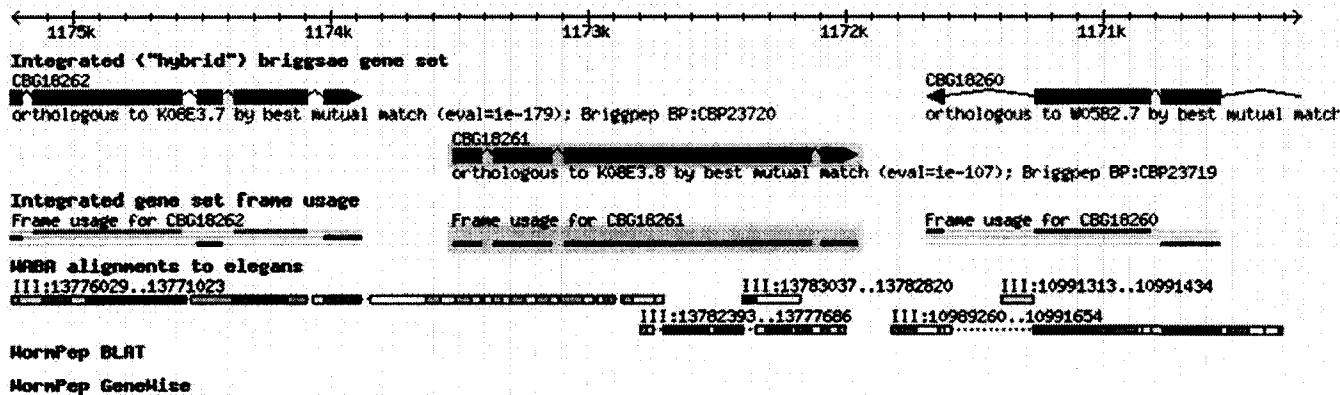
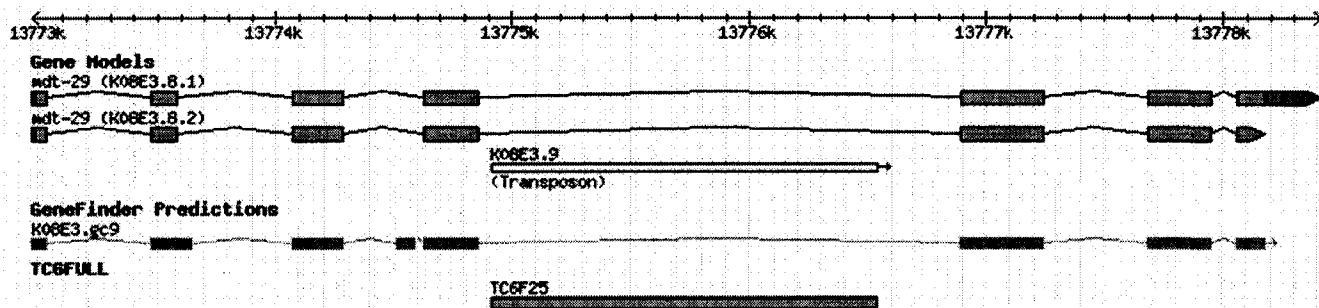
element matches 3253-6797, between exon 6&7 (2782-2932)(7526-7633)



130

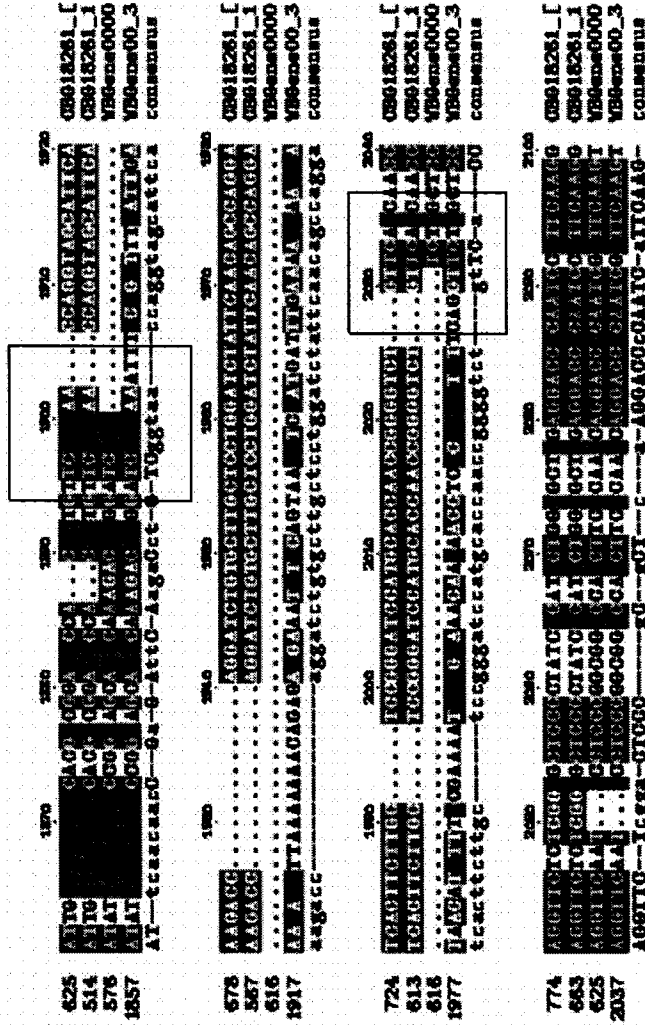


element matches 1945-3567, between exon 4&5 (1648-1878)(3910-4257)



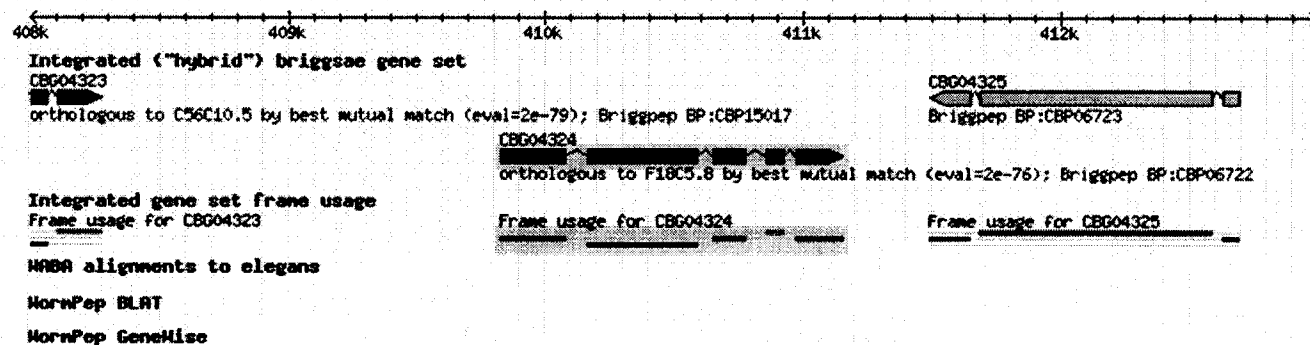
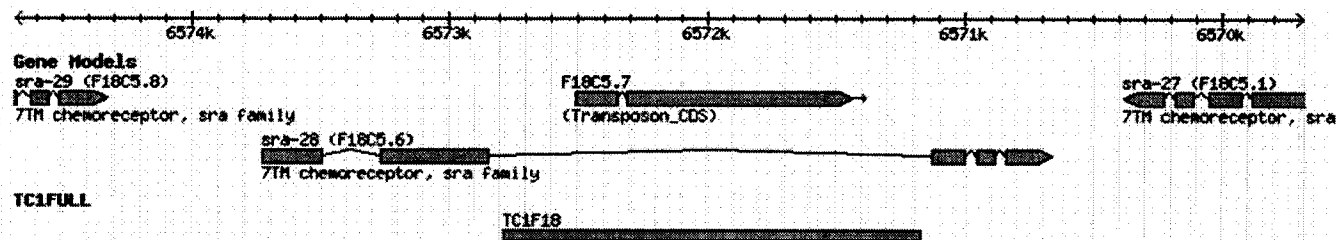
C elegans (cdna+(genomic-intron)/ C briggsae genomic and predicted cdna

b and e match, e potentially has longer intron due to element, unclear - briggs annotation defines as exon?

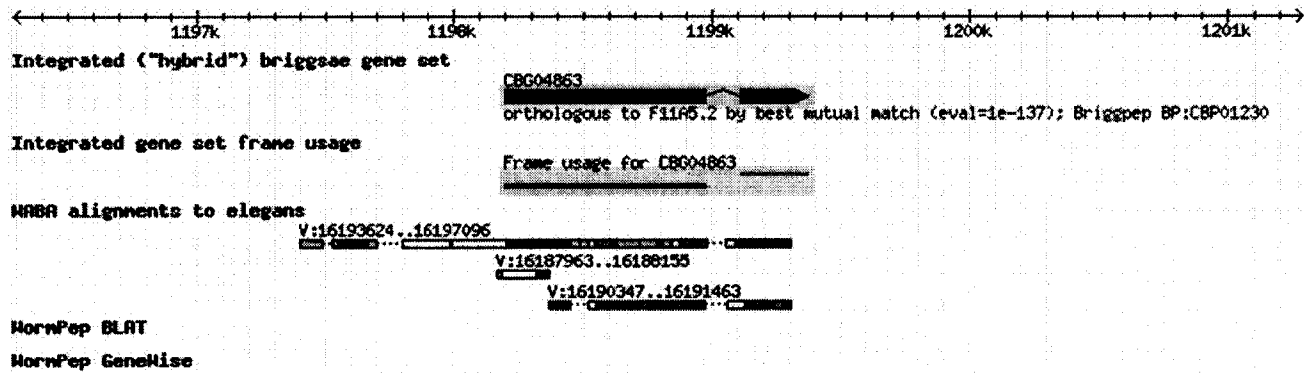
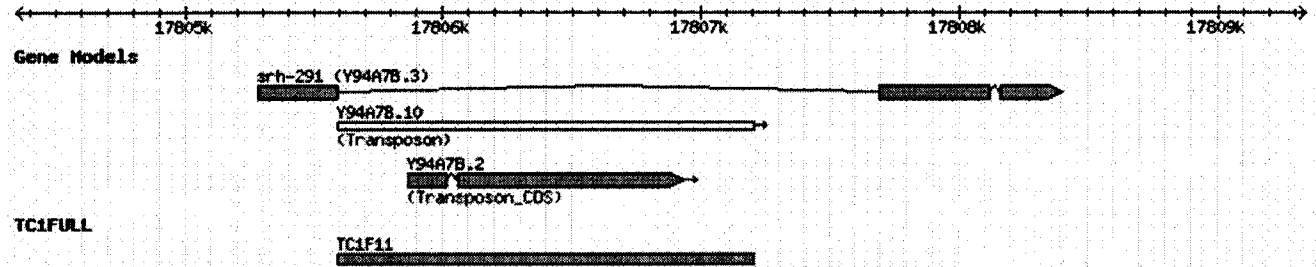


e gene much larger than b - longer introns as a whole

element matches 945-2555, between exon 2&3 (459-877)(2601-2725)



element matches 314-1924, between exon 1&2 (1-310)(2413-2837)



C elegans transcript

```

241 ATGGGAGTATTGGAATGGCTAAACGTGAATATGGGGATTATGGCGTTTTTCGGAATGATT
241 ATGGGAGTATTGGAATGGCTAAACGTGAATATGGGGATTATGGCGTTTTTCGGAATGATT
81 -M--G--V--L--E--W--L--N--V--N--M--G--I--M--A--F--F--G--M--I--
301 ATTGTTGCATTTCGTACCTATTTCCATCATCAAAATGTTGAAAACCGTTATTTTGTCTTG
301 ATTGTTGCATTTCGTACCTATTTCCATCATCAAAATGTTTGAACACCGTTATTTTGTCTTG
101 I-V-A-F-V-P-I-S-I-I-K-M-F-E-N-R-Y-F-V-L-
    
```

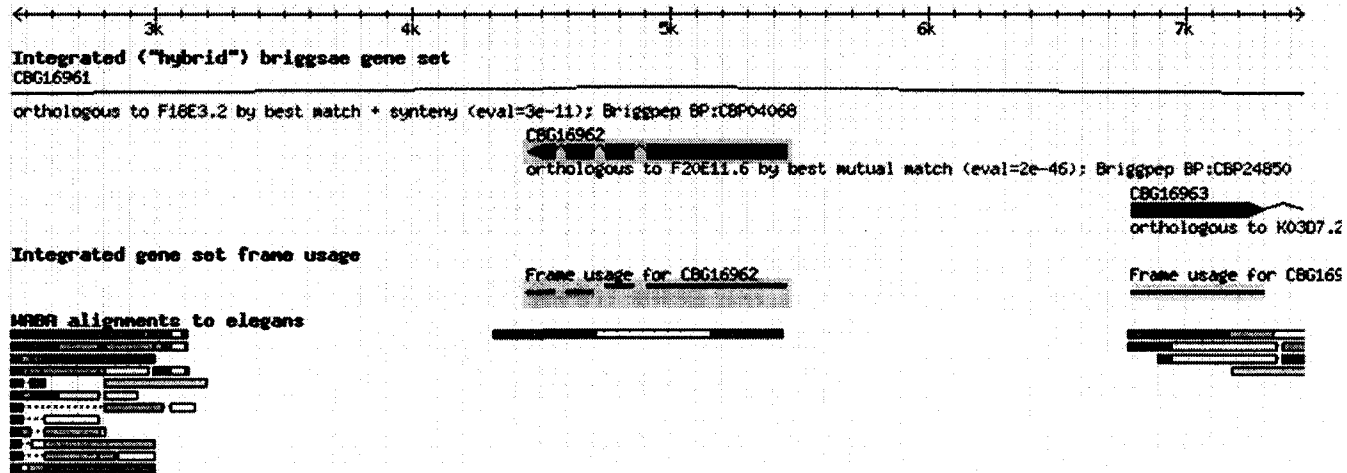
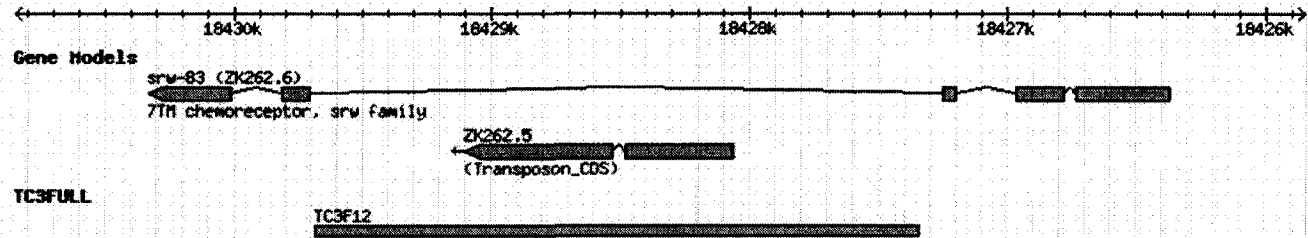
C elegans/C briggsae protein align.

	70	80	90	100	110	120	
60	YVSYFFYQ	CFPVKQ	CFK	EVLEIN	MA	YBGene0000
61	LLSLLTOP	KIVYS	IPY	YKFGYS	QS	MYSTLFCCECH	CB004863
	D-tvs	Qpyll-s	g-s	gil	Y-lal	yavstlfccschF-i	consensus
	130	140	150	160	170	180	
105	VP	SIIK	EEERYE	PANLTON	ETIYPR	ILST	YAST
121	VP	SIIK	EEERYE	PAENI	SRARYP	ATNY	NYLFP
	VavSII	iFENRYPLIYA	RTWR	RYPF	NYilalL	at-L-vP-Q-YAR-i-F	consensus
	200	210	220	230	240	250	
299	CGA	TTT	TTT	TTT	TTT	TTT	CB004863_1
299	CGA	TTT	TTT	TTT	TTT	TTT	CB004863_1
297	GAT	AAA	TTT	TTT	TTT	TTT	YBGene0000
297	GAT	AAA	TTT	TTT	TTT	TTT	Y94A7B.3
	t-TTgTT	caItgt	gagtgcc				consensus
	370	380	390	400	410	420	
323	CB004863_1
323	CB004863_1
357	YBGene0000
311	Y94A7B.3
	consensus
	430	440	450	460	470	480	
341	CB004863_1
341	CB004863_1
417	YBGene0000
311	Y94A7B.3
	consensus

b and e pretty good match, looks like briggs annotation, possibly e gene has larger intron due to element, but unclear

C elegans (cdna+(genomic-intron)/ C briggsae genomic and predicted cdna

element matches 977-3312, between exon 3&4 (832-877)(3335-3444)



C elegans transcript

```

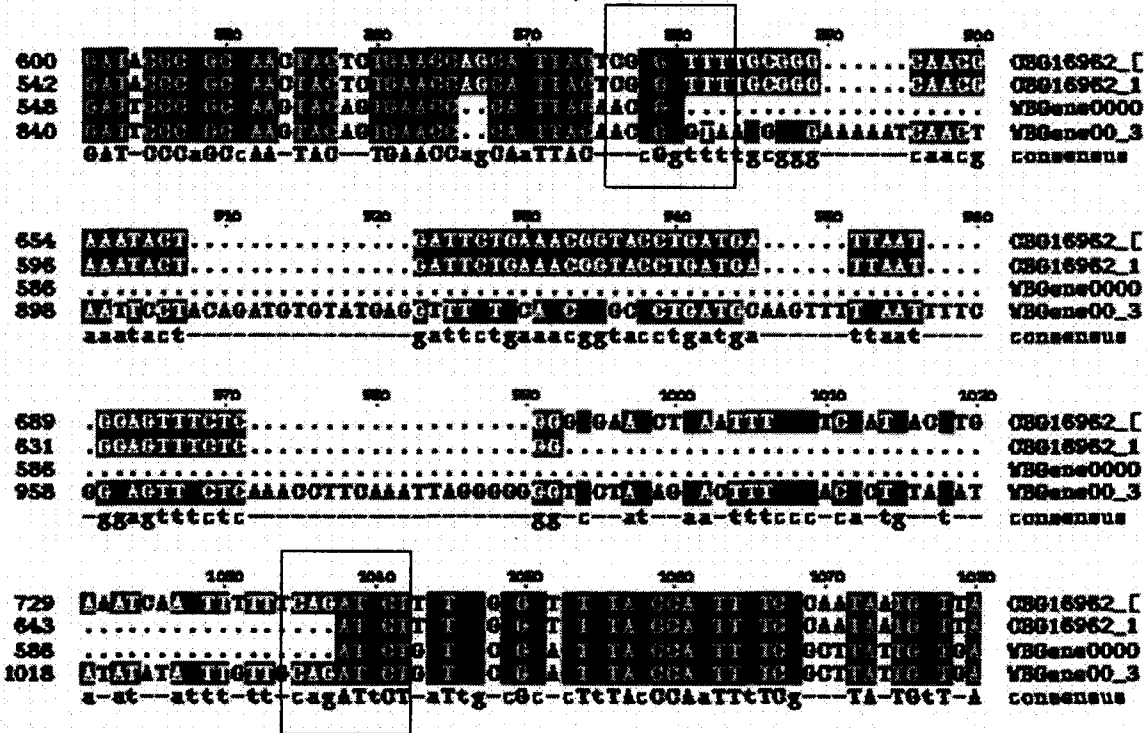
541 TGCACCGGATTCCCAGCCAAGTACAGTGAACCCAATTACAACCGGATTCTGATTGCCGCA
541 TGCACCGGATTCCCAGCCAAGTACAGTGAACCCAATTACAACCGGATTCTGATTGCCGCA
181 -C--T--G--F--P--A--K--Y--S--E--P--N--Y--N--R--I--L--I--A--A--
601 CTTTACCCAATTTTCGGCTTATTGTTGATGTTTGAAGTGTGAAAGCAGCCAAAGTTGCG
601 CTTTACCCAATTTTCGGCTTATTGTTGATGTTTGAAGTGTGAAAGCAGCCAAAGTTGCG
201 -L--Y--P--I--F--G--L--L--L--M--F--E--V--L--K--A--A--K--V--A--
    
```

C elegans/C briggsae protein align.

```

          180          200          220          240
179 AECYQ PAKTSEP.....NYEIL AALYVFG LSE LAA YBGene0000
177 EECYQ PANYSEPAYTEDEAGEEILIKYILIEGVSRIL SVFYELT LSY EIST CBG16962
    --GTGyPA-YSEPavtrdfagneililkrylwin--RILv--YFv--imLIF-i-K-- consensus
    
```

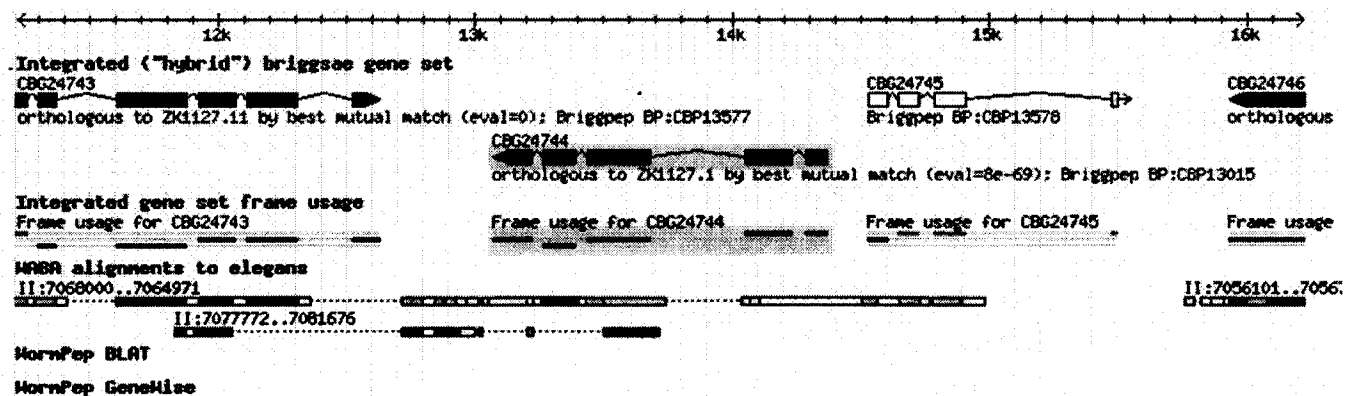
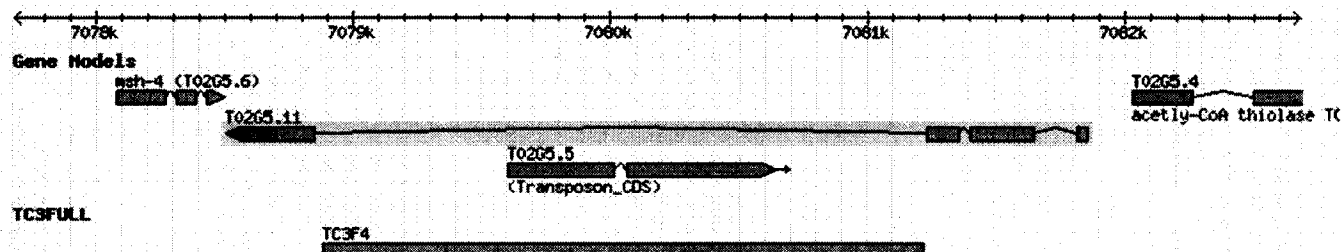
C elegans (cdna+(genomic-intron)/ C briggsae genomic and predicted cdna



141

briggs exon extends into where e element intron begins, likely that both have intron (annotation issue), and e has longer due to element

element matches 633-2969, element is on other strand



C elegans transcript

```

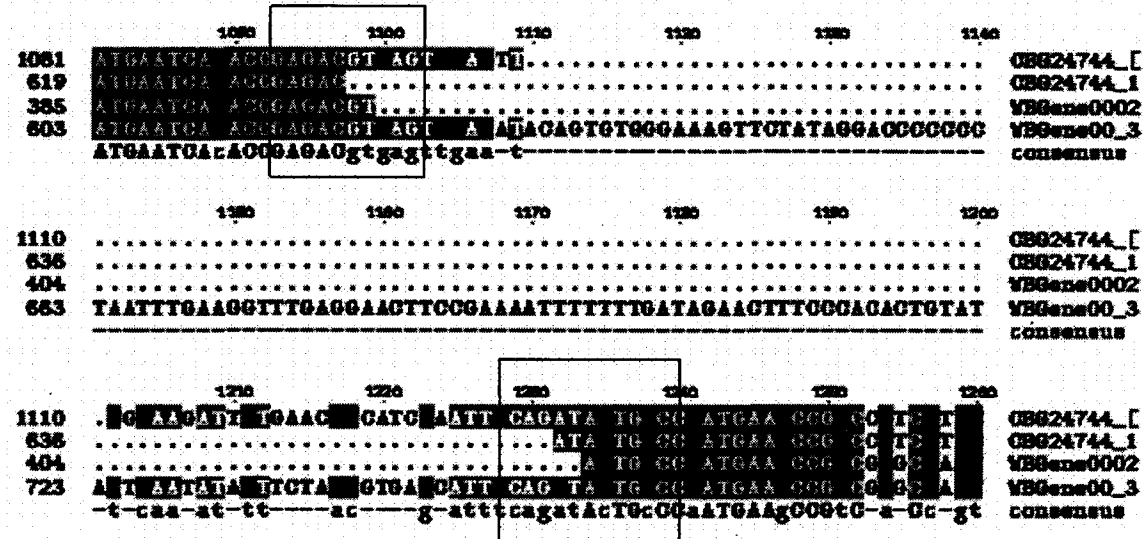
361 AAAATTGTTGGAGCTCGGGGTGAAATGAATCACACCGAGACGACTGCCAAATGAAGCGCG
361 AAAATTGTTGGAGCTCGGGGTGAAATGAATCACACCGAGACGACTGCCAAATGAAGCGCG
121 -K-I--C--G--A--R--G--E--M--N--H--T--E--T--Y--C--P--M--K--P--
421 TCGAGCCAGTTGTCCTCAATGAGGATTCAGCCGAGATTCGAAAACCGCCGATTCAG
421 TCGAGCCAGTTGTCCTCAATGAGGATTCAGCCGAGATTCGAAAACCGCCGATTCAG
141 -S--S--Q--L--S--F--N--E--D--F--S--R--D--F--E--N--R--R--F--Q--

```

C elegans/C briggsae protein align.

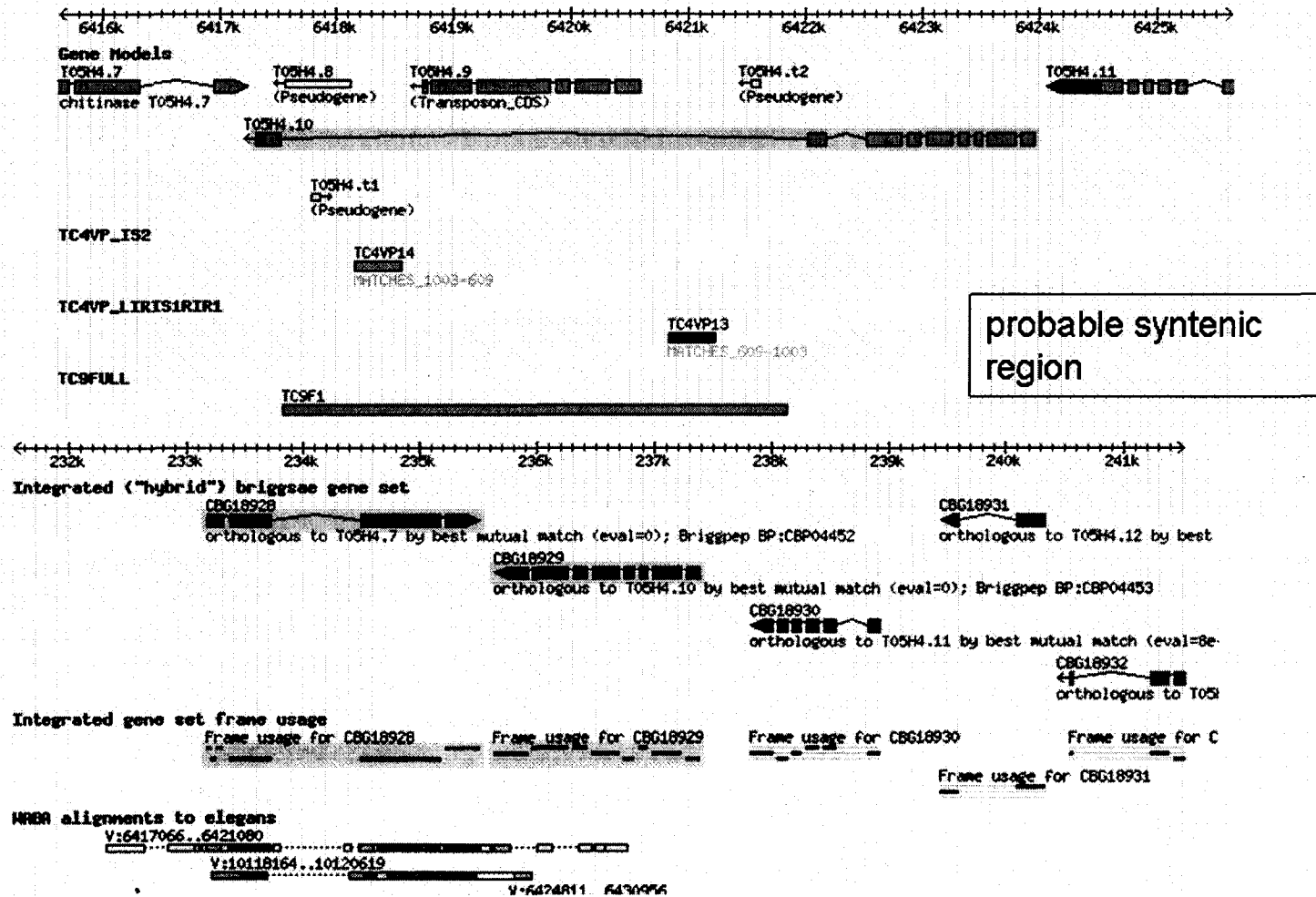
1	KL	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300
1	MS	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300
15	LE	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300
43	Y	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300
103	VE	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300
131	VE	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300

C elegans (cdna+(genomic-intron)/ C briggsae genomic and predicted cdna



b and e both match, e has larger intron due to element

element matches 2124-6416, on other strand, between exon 8&9 (1790-1939)(6440-6580), whole briggs genomic alignment 1-1787 of elegans gene, both proteins align very well (476 aa)



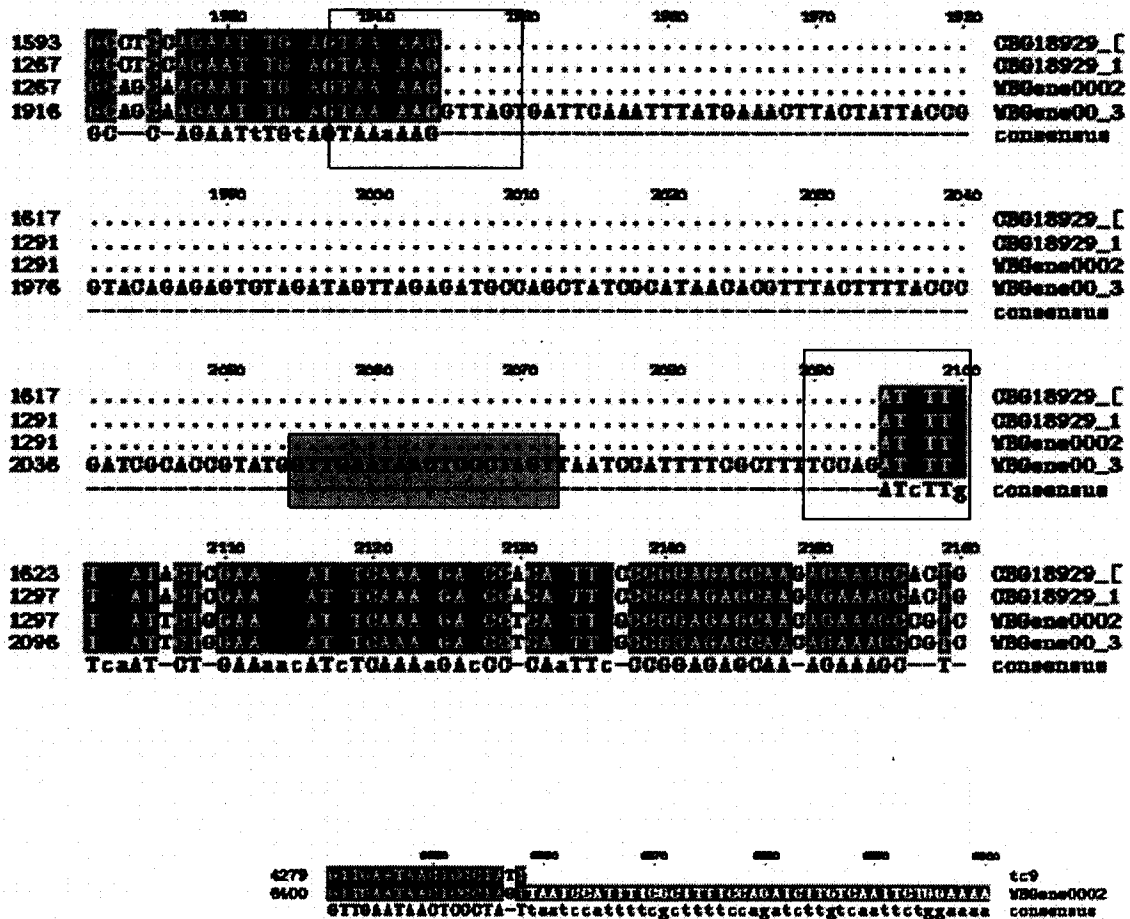
C elegans transcript

1261 ACTTCTGCAGCAAGAATTTGTAGTAAATAAGATCTTGTCAATTCCTGGAAAAACATCTCAAAA
 1261 ACTTCTGCAGCAAGAATTTGTAGTAAATAAGATCTTGTCAATTCCTGGAAAAACATCTCAAAA
 421 -T--S--A--A--R--I--C--S--K--K--I--L--S--I--L--E--N--I--S--K--
 1321 GACCCCTCAATTGGCCGAGAGCAACAGAAAGCCGTCAATTCAATTGAGAAAAATTGAAAAG
 1321 GACCCCTCAATTGGCCGAGAGCAACAGAAAGCCGTCAATTCAATTGAGAAAAATTGAAAAG
 441 -D--P--Q--F--A--G--E--Q--K--A--V--I--S--I--E--K--I--E--K--

C elegans/C briggsae protein align.

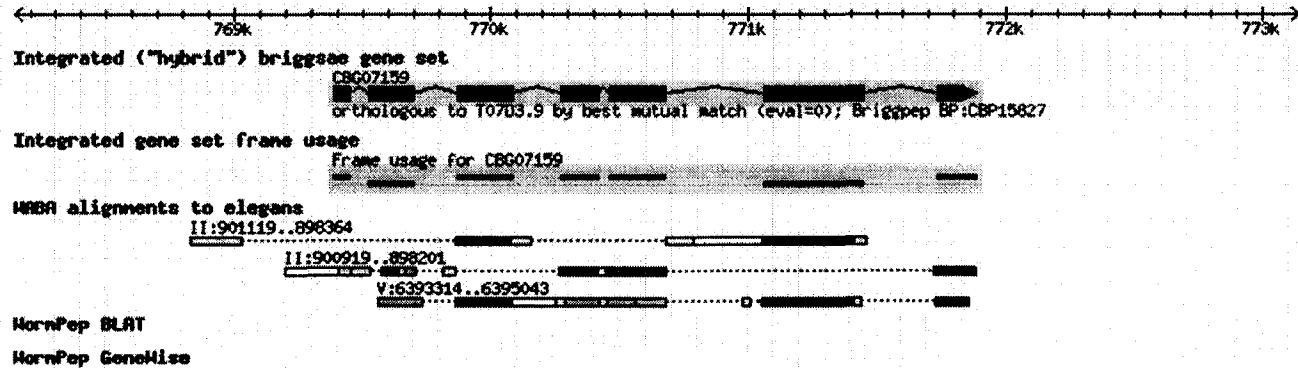
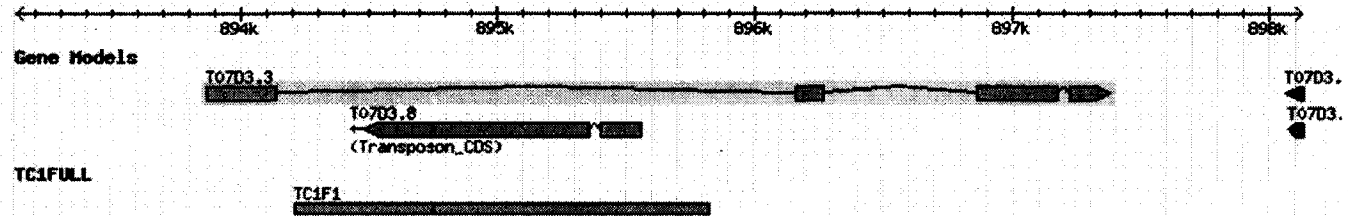
361 VADPTLVGQTRAVLSSVTSRGRSVAQNTLAQYFSSLSSESRVQRGKGEQACKAA VBQsae0002
 361 VADPTLVGQTRAVLSSVTSRGRSVAQNTLAQYFSSLSSESRVQRGKGEQACKAA CB018929
 VADLFTREGLEAWI-SVIESGEPFRVETGTYLAQYFCSLLEFEERVQRGKGEQACKAA consensus
 421 TSAHIGSEKILSLESDPQFGEQCKAISLIEKAIQEKYKPKKPKSSS VBQsae0002
 421 TSAHIGSEKILSLESDPQFGEQCKAASLIEKAIQEKYKPKKPKSSS CB018929
 T-A-RIGSKLL-ILE-ISEDPQFGEQ-IAL-SI-TIEKAIQKMKKIKKFI-EZ consensus

C elegans (cdna+(genomic-intron)/ C briggsae genomic and predicted cdna

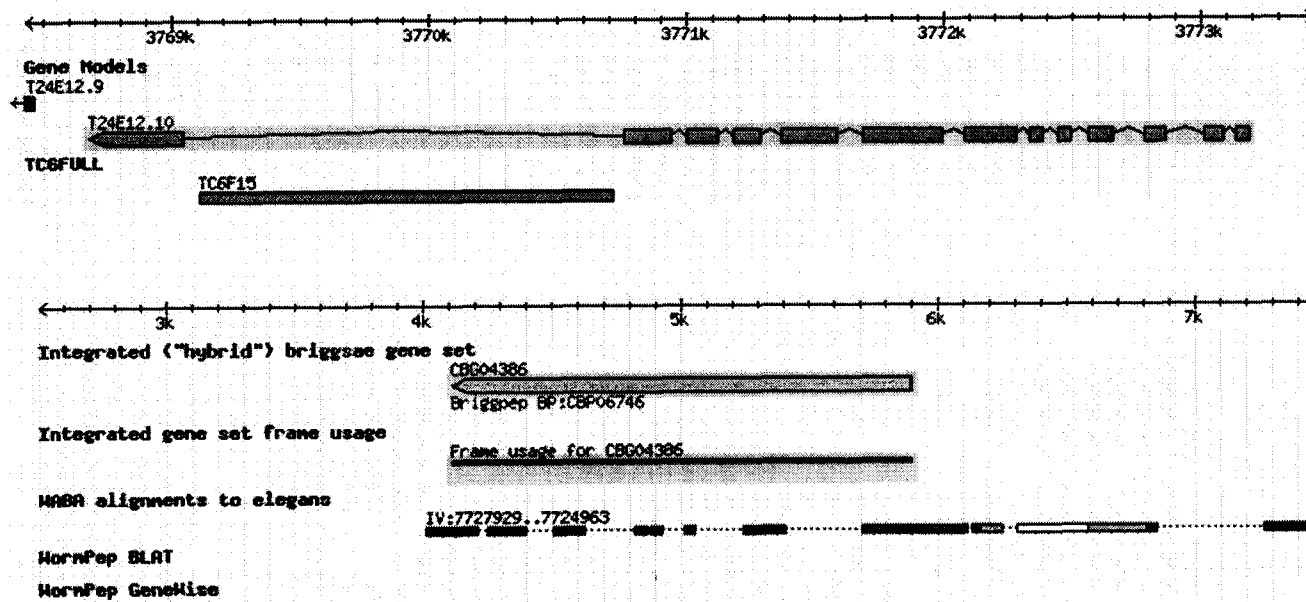


element created intron in e gene – b does not have an intron but aligns very well to surrounding region

element matches 349-1959, between exon 1&2 (1-273)(2298-2402), protein alignment not that good (briggs protein (448aa) elegans (284aa),



element matches 2473-4076, between exon 12&13 (2249-2428)(4142-4509), briggs and elegans protein not good alignment, but genomic alignment good – intron, no briggsae introns predicted in sequence available



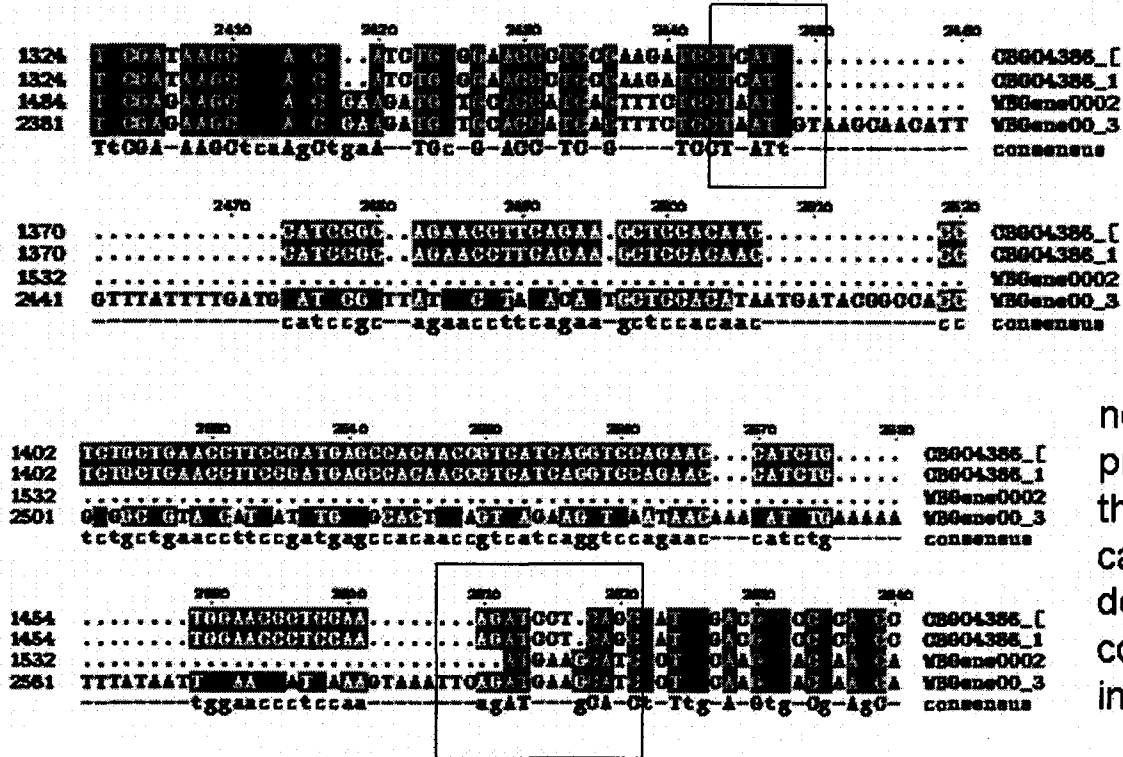
C elegans transcript

1501 GAAGATGCTGCACCACATCAGTTTCTCTAATATGAGCATCTCTTGGCAAGTGCAGCAAGCA
 1501 GAAGATGCTGCACCACATCAGTTTCTCTAATATGAGCATCTCTTGGCAAGTGCAGCAAGCA
 501 -E--D--A--A--P--S--V--S--P--N--Y--E--A--S--L--A--S--D--E--A--
 1561 CTTTGTGAGATTGTTGTGGAGTCAAGAAGATCGCCAATCTCATCTCTCCAGCTAGC
 1561 CTTTGTGAGATTGTTGTGGAGTCAAGAAGATCGCCAATCTCATCTCTCCAGCTAGC
 521 -L--C--E--I--V--V--E--V--Q--E--D--R--Q--F--S--S--P--A--S--

C elegans/C briggsae protein align.

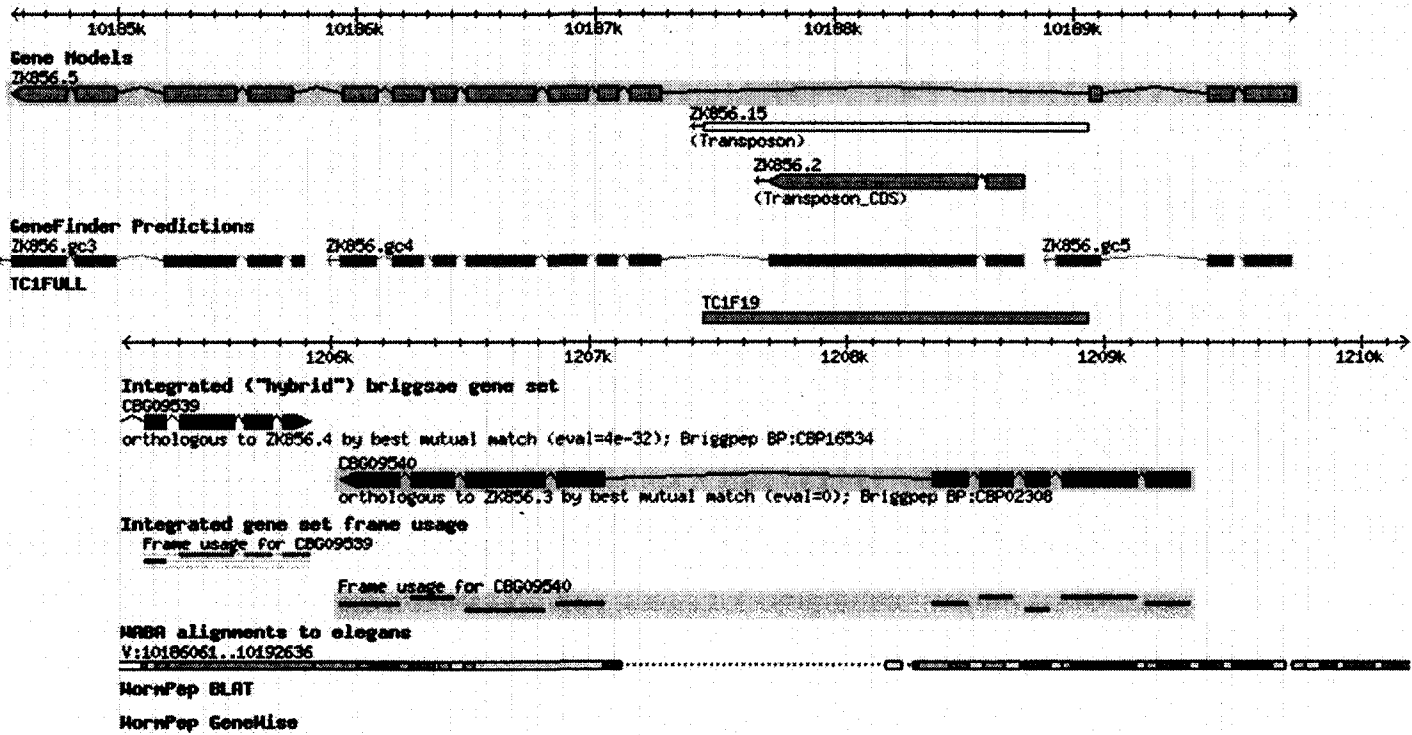
471	YV	EV	EL	GL	VV	AV	YS	YS	AS	AS	YS	DE	YBGenes0002
486	SE	P	E	L	P	A	E	S	P	Q	S	D	PS
	---	sg	ub	---	g	v	p	z	---	q	us	---	vr
531	SS	AR	HM	AI	KS	Q	S	E	EM	VI	IP	SA	HL
480	FP	DA	OP	AE	SS	EL	EP	SA	EP	S	K	Q	P
	Dr	Q	S	---	P	---	u	---	e	---	q	u	---
590	AE	HE	..	NI	PE	PE	..	PS	VQ	AM	EIE	DS	Q
550	KE	PP	EM	PE	PE	GP	GP	TR	PP	...	HE	KGV	VFD
	---	e	---	d	---	r	---	q	---	v	---	g	---

C elegans (cdna+(genomic-intron)/ C briggsae genomic and predicted cdna



no b introns predicted in this ortholog, can not determine conservation of intron

element matches 866-2475, between exon 3&4 (804-849)(2643-2775), good protein alignment (first 181 aa of e gene not match briggs, but 181-760 matches whole briggs protein (563aa) well



C elegans transcript

```

301 TCACATTGATGGATATGGATGGGAAATTGTTCTTGATGCCGTCAATCAAACCACCTCGCT
288 TCACATTGATGGATATGGATGGGAAATTGTTCTTGATGCCGTCAATCAAACCACCTCGCT
96 --H--I--D--G--Y--G--W--E--I--V--L--D--A--V--N--Q--T--T--S--L

361 TCTTTTCTCCAGATTCCCAAATCAAACAATTCTTCCAACATTTGGAAATCAGGACTACGC
348 TCTTTTCTCCAGATTCCCAAATCAAACAATTCTTCCAACATTTGGAAATCAGGACTACGC
116 --L--F--S--R--F--P--N--Q--T--I--L--P--T--F--G--N--H--D--Y--A
    
```

<p>61 1</p>	<p>70 80 90 100 110 120</p> <p>DINCDAPEPLIQLAIDESAIEPEPDLIIYIGDEVAEIDGTGVEIWLDAVEOTLSRASE</p>	<p>WBGene0001 CB009540 consensus</p>	<p>C elegans/C briggsae protein align.</p>
<p>121 1</p>	<p>130 140 150 160 170 180</p> <p>FPRQTLRSTGQEDLAPSSGFESESSLIETVELYEGYLODESKATFLEGGYATRLSEA</p>	<p>WBGene0001 CB009540 consensus</p>	
<p>181 1</p>	<p>190 200 210 220 230 240</p> <p>TAVVLETERAASKAYVEFIEQVADQAFLEKELSAKQPRKSEEDISYRIAR</p>	<p>WBGene0001 CB009540 consensus</p>	

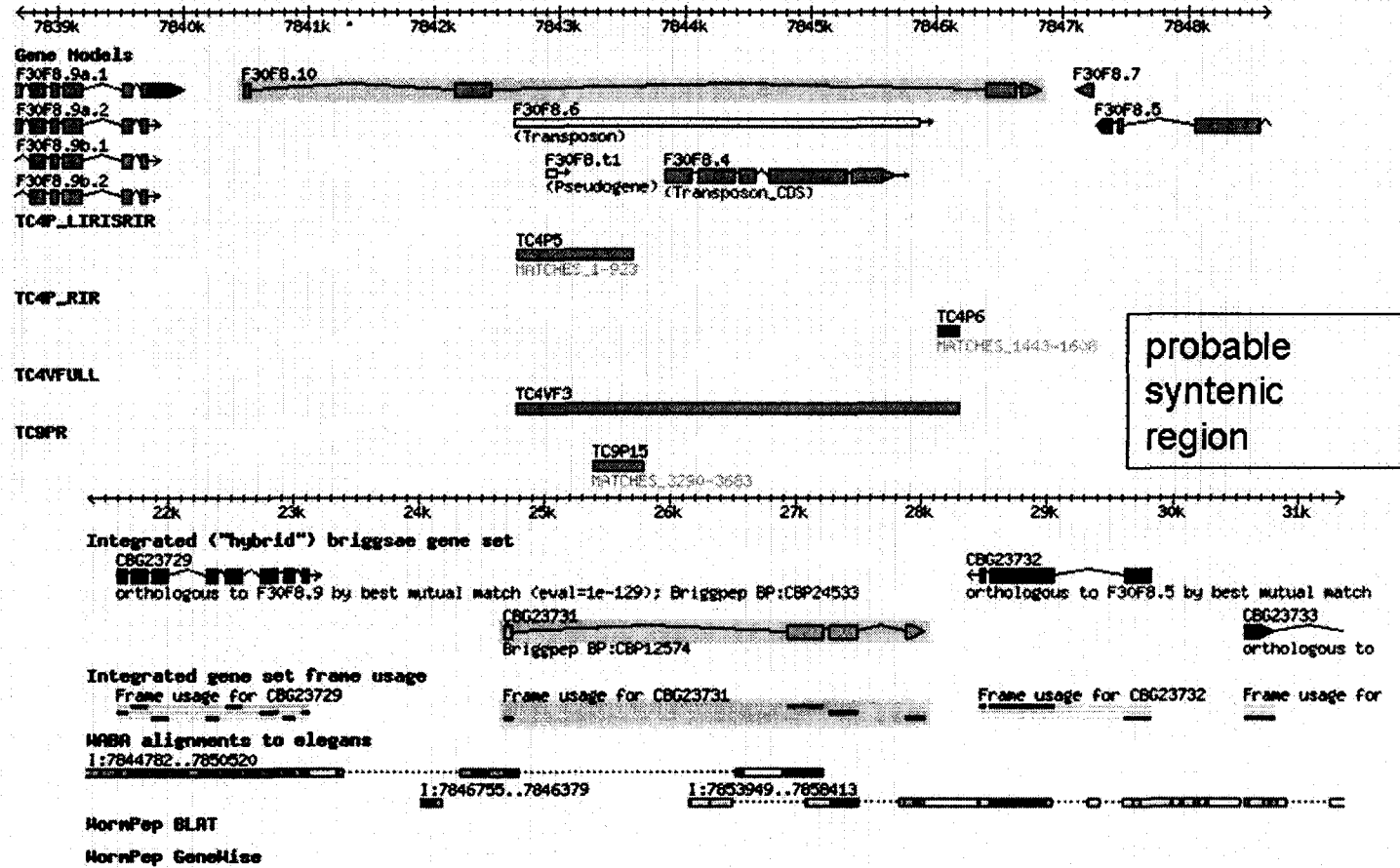
154

C elegans (cdna+(genomic-intron)/ C briggsae genomic and predicted cdna

briggs ortholog matches after intron where element is

<p>1 351 841</p>	<p>CGAGCTCCGCTGCTTTCTCGAG.....</p> <p>CGAGCTCCGCTGCTTTCTCGAGCTAGAGTCTGCGCAAAAAGATATCGACTTTTGCTTT</p> <p>ccactctgcttctttctccag</p>	<p>CB009540_1 CB009540_1 WBGene0001 WBGene00_3 consensus</p>
<p>1 1 373 901</p>	<p>TTGTGTATACTTTTGTGAAAGATGCAATTGACTGAAATTTTGTGTTGTAAGAAAAG</p>	<p>CB009540_1 CB009540_1 WBGene0001 WBGene00_3 consensus</p>
<p>1 373 951</p>	<p>TTGCGCAAAATCAAAAC</p> <p>TTGCGCAAAATCAAAAC</p> <p>tttccccactcaaac</p>	<p>CB009540_1 CB009540_1 WBGene0001 WBGene00_3 consensus</p>

element matches 2197-5714, between exon 2&3 (1696-1980)(5919-6151),



C elegans transcript

301 TCAGAAACCAAGTTCCTCTATGATGTAATCTTGAGGATCAAGGAATCCTGCCATCTTCA
 301 TCAGAAACCAAGTTCCTCTATGATGTAATCTTGAGGATCAAGGAATCCTGCCATCTTCA
 101 -S--E--T--K--F--L--Y--D--C--N--I--E--D--Q--G--I--L--P--S--S--
 361 AACGCCACGTCGCCATCATCTATCAATAGCTGTGATAAAAAATCCGAAGATTAGGT
 361 AACGCCACGTCGCCATCATCTATCAATAGCTGTGATAAAAAATCCGAAGATTAGGT
 121 -N--A--H--V--A--Y--I--L--S--I--A--V--D--K--K--F--R--L--G--

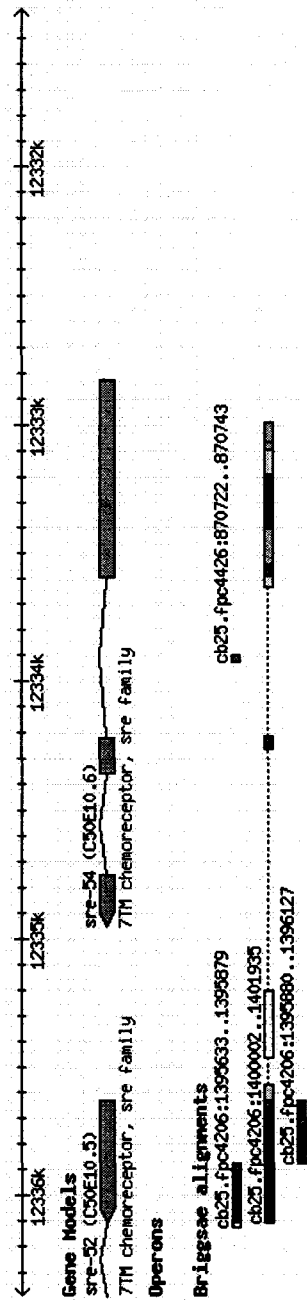
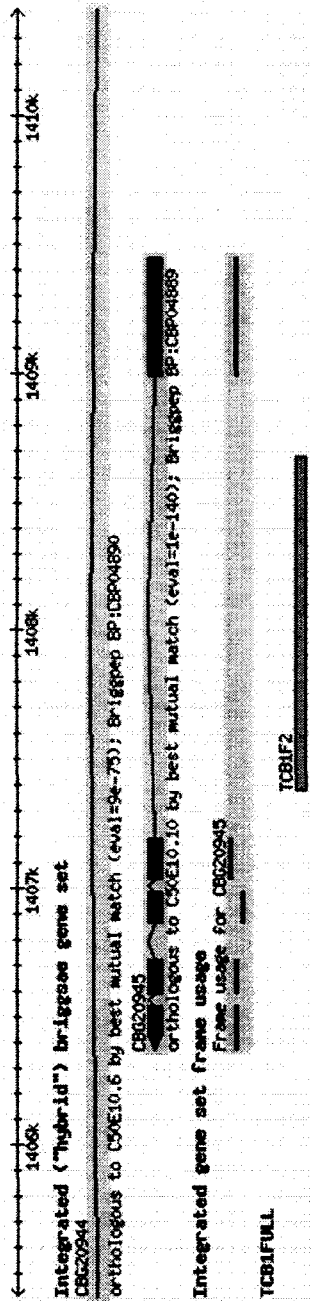
C elegans/C briggsae protein align.

1	MEQ...	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240
1	MTESEHSVSHGYSPISSS	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240
55	RVAEALGKESDPQIRQVMDYVSSGLVHGRNCEAAKVSEIKFV	2621	3300	3350	3400	3450	3500	3550	3600	3650	3700	3750	3800	3850	3900	3950	4000	4050	4100
61	RAVE-LOHESFPQIPDINWDEVHVSQGL-STGLFDGEnLAANIYSETEPv-DEGLEDD-	315	1868	1968	2068	2168	2268	2368	2468	2568	2668	2768	2868	2968	3068	3168	3268	3368	3468

C elegans
 (cdna+genomic
 -intron)/C
 briggsae
 genomic and
 predicted cdna

e and b both have intron, e has longer intron due to element

C.briggsae Intron Study
(B Intron Study)



Exon #	Start	End
1	2905	3076
2	2727	2859
3	2466	2583
4	2259	2415
5	1	458

briggsae transcript and protein alignment using Wise2

```

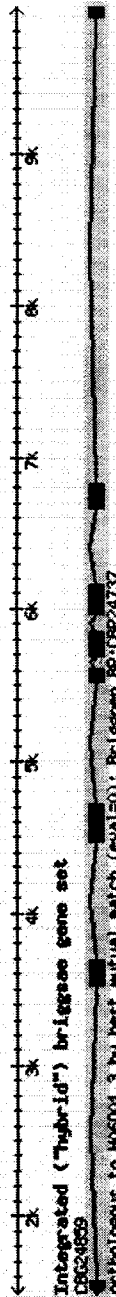
CBG20945      148 ATYFI      DYERTPRTYIGVGLLFSV
                ATYFI      DYERTPRTYIGVGLLFSV
                ATYFI      R:R[agg]      DYERTPRTYIGVGLLFSV
CBG20945      442 gattaAGGTACAGT Intron 1 CAGGgtgcaccatagggccttg
                ccatt <2-----[459 : 2258]-2> aaagccgcatgtgtttct
                ggttc      ccgaaaactcagttttgt
    
```

159

```

                120      140      160      180      200
115 FTI PLF PG NI ET YS FI TC VS ER CA YI RD ET ET TV GV GC FS VHC CBG20945
120 GEA PLF PG NI ET YS FI TC VS ER CA YI RD ET ET TV GV GC FS VHC OSOE10.6
                PLFvgGvL-WyTs-K-gi-vVSLERICA-TFI-VE-PL-VI-Li-qfl consensus
    
```

briggsae and elegans protein alignment



Sytemy block

Contigs and clones
c001500106.Contig3

cb25.Nr_236

TCB1FULL
TCB1F6

Exon #	Start	End
1	8401	8480
2	6273	6439
3	5245	5492
4	4365	4442
5	4118	4275
6	3811	3996
7	3139	3301
8	1	62



Gene Models

H6001.2

chromodomain-helicase-DNA-binding protein H6001.2

ctg-1 (H6001.3)

retinal-binding protein like

H65001.4

protein-tyrosine phosphatase

Operons

Briggsae alignments

cb25.fpc2374:448669..450955

cb25.fpc2374:450956..452373

cb25.Nr_238:3961..6829

cb25.Nr_238:3522..3699

cb25.fpc2374:315033..315204

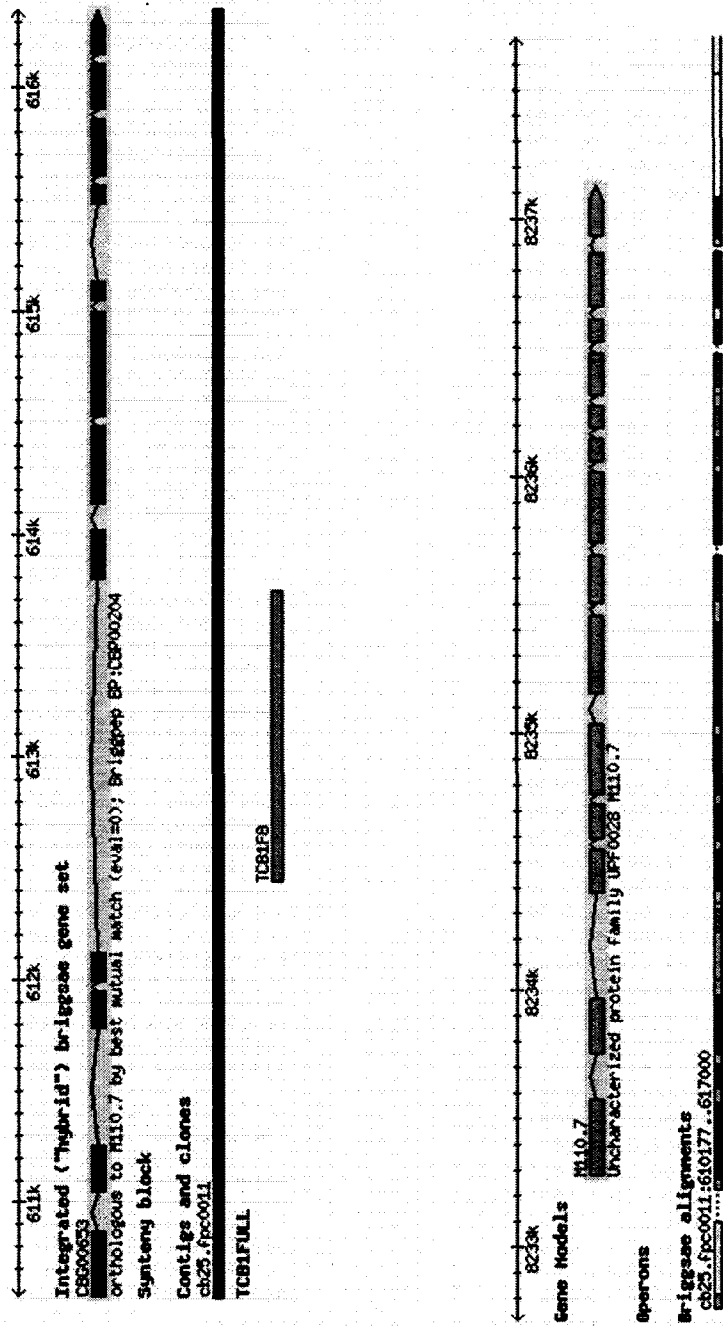
cb25.fpc4184:193726..194149

cb25.fpc1402:802182..

cb25.Nr_018:88..690

cb25.Nr_220:10503..1

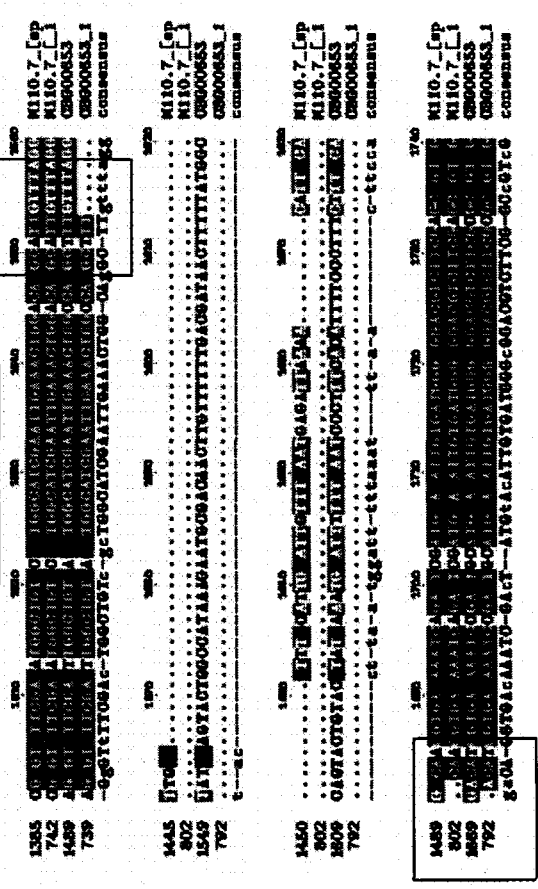
cb25.fpc1402:180

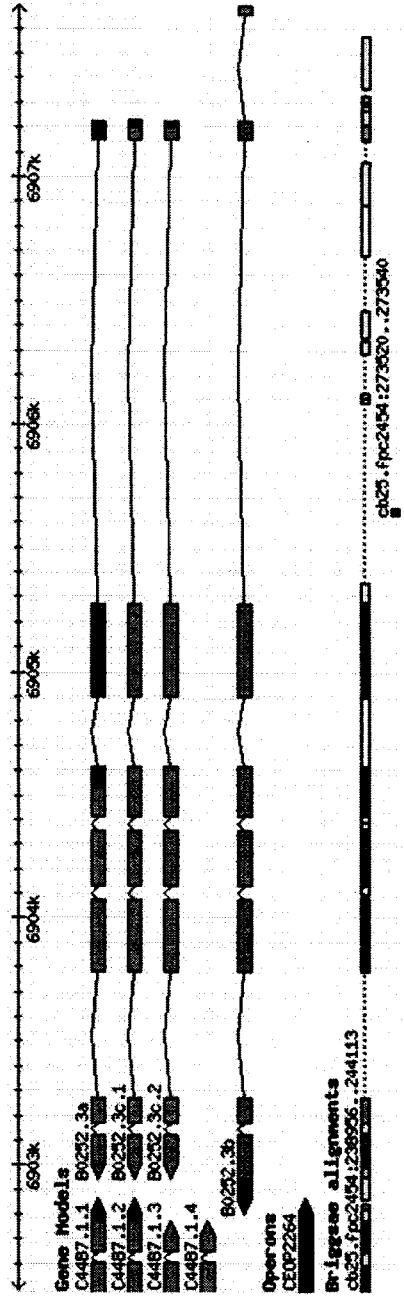
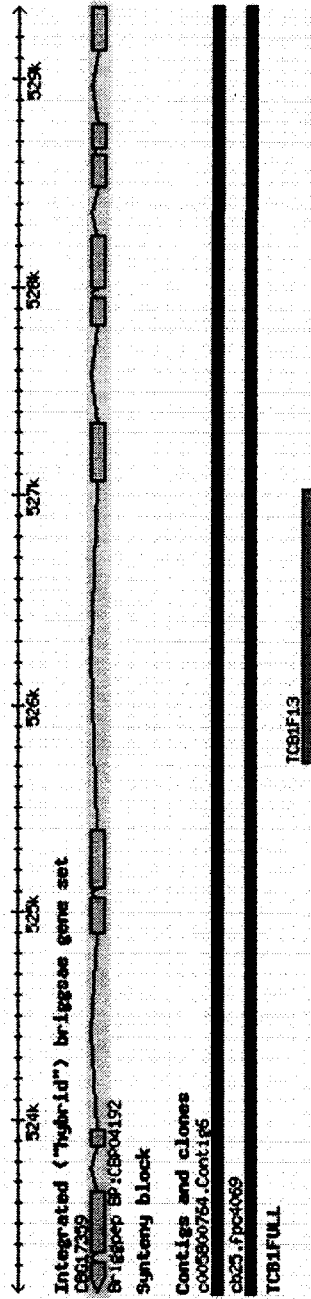


CBG00653 246 ARVFDLAVSWHRIETGQAL- HGD
 ARVFDLAVSWHRIETGQAL HGD
 ARVFDLAVSWHRIETGQALF HGD
 CBG00653 1486 gagtggatccagagcgttaggtataca --:R[aga] Intron 4 CAGAcgg
 cgttatctggagtagactt <2-----[1548 : 3219]-2> aga
 cgtccgtccgtatcgtg ttc

240 HE SPFAKYD AV VHRLEIGQAL..EQDSDQV IYGGALSAVDYIEIIEVGEELDI CBG00653
 241 F SPFAKYD AV VHRLEIGQALFR QXSD FHLVGGELLEAVDSIKIIEVGEELDI M110.7
 qFISPFARVFD~~AV VHLIEIGQALfrqQVSD~~~~D~~~~NY IYVGGELLEAVD~~~~IKIIEVGEELDI~~ consensus

b intron begins
 before e does,
 but next exon is
 same






```

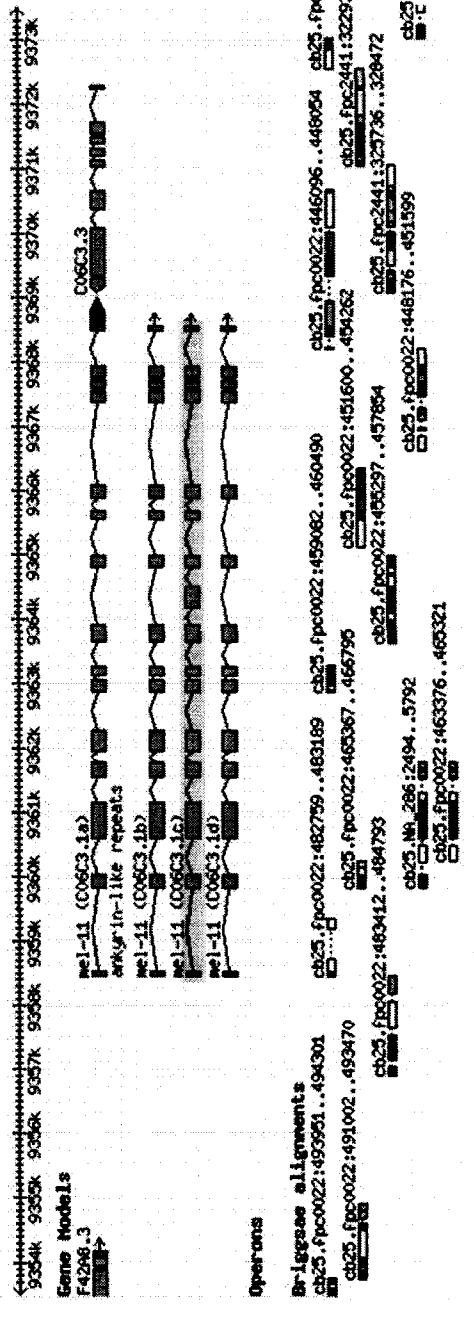
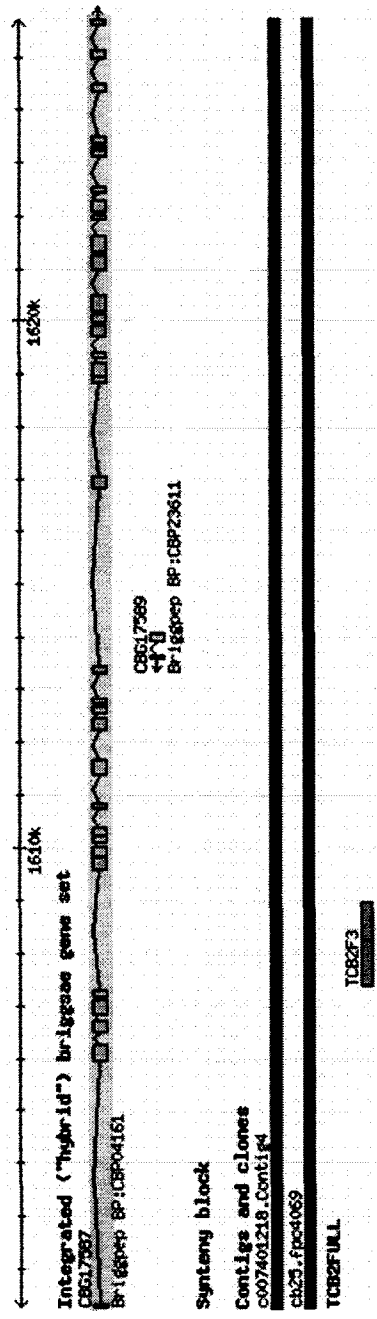
341      2740      2750      2760      2770      2780      2790      2800      2810      2820      2830      2840      2850
G A T C A T T C A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C
2246      2740      2750      2760      2770      2780      2790      2800      2810      2820      2830      2840      2850
C A T C A T T C A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C
2253      2740      2750      2760      2770      2780      2790      2800      2810      2820      2830      2840      2850
C A T C A T T C A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C
1117      2740      2750      2760      2770      2780      2790      2800      2810      2820      2830      2840      2850
c c a a g t a t t c a c a t t t a t t a t c a t t a t g a t g a t t t t g t a t a a t g a g a a t g a t t c
consensus

401      2860      2870      2880      2890      2900      2910      2920      2930      2940      2950      2960      2970      2980      2990      3000
A T C G G . A T A T T T A T A C A G . . . . . A A C A T C A T T A T G A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C
2308      2860      2870      2880      2890      2900      2910      2920      2930      2940      2950      2960      2970      2980      2990      3000
A T C G G . A T A T T T A T A C A G . . . . . A A C A T C A T T A T G A T T C A G A T T C A G A T T C A G A T T C A G A T T C A G A T T C
2312      2860      2870      2880      2890      2900      2910      2920      2930      2940      2950      2960      2970      2980      2990      3000
A T C G A A A A A T A C T C C C T A A A A T G C A C A A A T T A C A A A A A C A A A T T C G A A I
1141      2860      2870      2880      2890      2900      2910      2920      2930      2940      2950      2960      2970      2980      2990      3000
a t g g g - a t a t t t t a t a c a g - a t - g - a - c - a - a a c c a a - a - a c a - g - - t c g c - t t
consensus

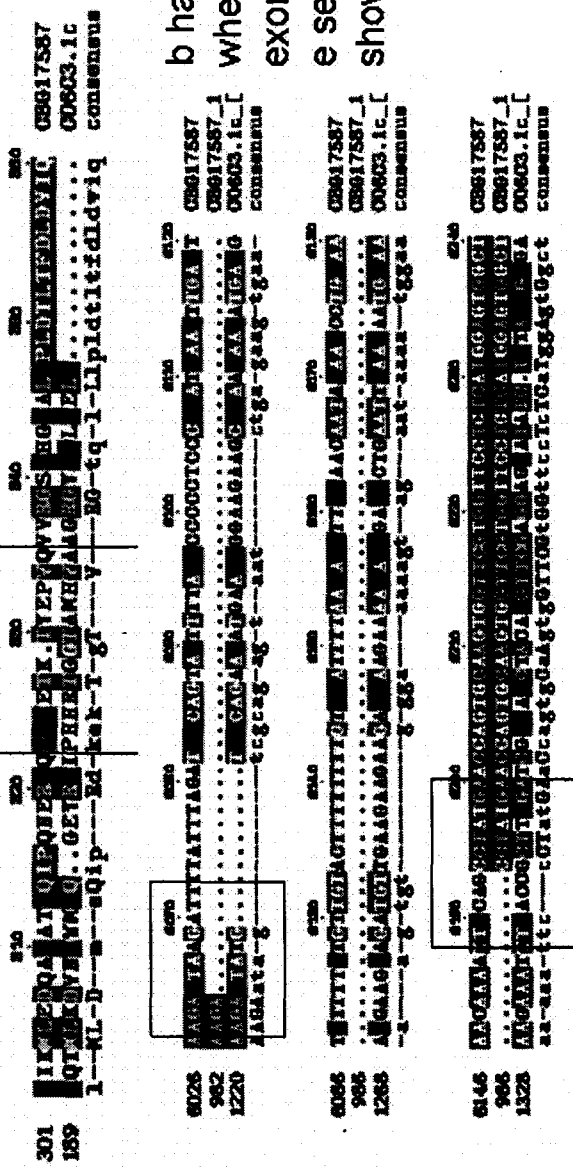
419      3010      3020      3030      3040      3050      3060      3070      3080      3090      3100      3110      3120      3130      3140      3150
. . . . . A A A G T T C A G C A T T C G A A A A A C A T T C C A G T C A C A G A T A T C . . . . .
2365      3010      3020      3030      3040      3050      3060      3070      3080      3090      3100      3110      3120      3130      3140      3150
. . . . . A A A G T T C A G C A T T C G A A A A A C A T T C C A G T C A C A G A T A T C . . . . .
2372      3010      3020      3030      3040      3050      3060      3070      3080      3090      3100      3110      3120      3130      3140      3150
C T T T A T C A T T C G A A A A A C A T T C C A G T C A C A G A T A T C . . . . .
1141      3010      3020      3030      3040      3050      3060      3070      3080      3090      3100      3110      3120      3130      3140      3150
. . . . . t g g c c a g - a c t - t - a - a - c t - a - c t a t t g t g t a t t c c t t c c t t c a t a t t c t g t a g
consensus

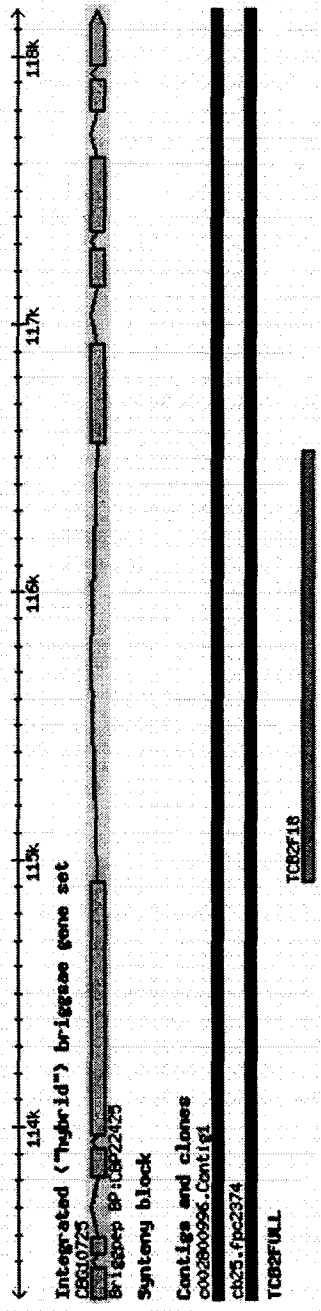
```

b and e not good protein alignments,
and genomic b exon where e still intron



CBG17587 289 DNLELFHRSRTMIKMLEDDQAIATQIPQNEREQKEKETK
 DNLELFHRSRTMIKMLEDDQAIATQIPQNEREQKEKETK
 DNLELFHRSRTMIKMLEDDQAIATQIPQNEREQKEKETK
 CBG17587 5909 gatgctccacaaaaacggcgagaaacaccagcgagagaa
 aatattagaggcttataaactcccatcaaaagaaaaaca
 ccgagccgttatcggggtatccctgtggtgaaaaaaagg
 CBG17587 329 YEPVQVVRGSSHGVALIPLDTLT
 YEPVQVVRGSSHGVALIPLDTLT
 YEPVQVVRGSSHGVALIPLDTLT
 T:T[acc]
 CBG17587 6029 AGTAAGAT Intron 5 CAGCCTgcgcggttcgggaccggaca
 <1-----[6030 : 8308]-1>
 aactattggccagtccttctactc
 taagagtttcttagttcatcagc





CBG10725 404 RRPAND ANHSEIFDSQQPKQYCE
 RRPAND ANHSEIFDSQQPKQYCE
 RRPAND ANHSEIFDSQQPKQYCE
 CBG10725 1536 cccgagAGGTACTGG Intron 4 CAGTgactgatgacccacttg
 ggccaa <2-----[1556 : 3194]-2> caacattagaacaaaaga
 atactc atcgaatttgagaactg

S 1500 1520 1540 1560 1580 1600
 1501 **TTTTCAGACAAAGTATTCAGACGCGAACAATCGGACGCGGACCGGAAATGACAGGTAAGI** tcb2_m6430
 tttccacagagcttatggagatggaaactgssgagcggctccagccaatgataggtaGI consensus
 CBG10725

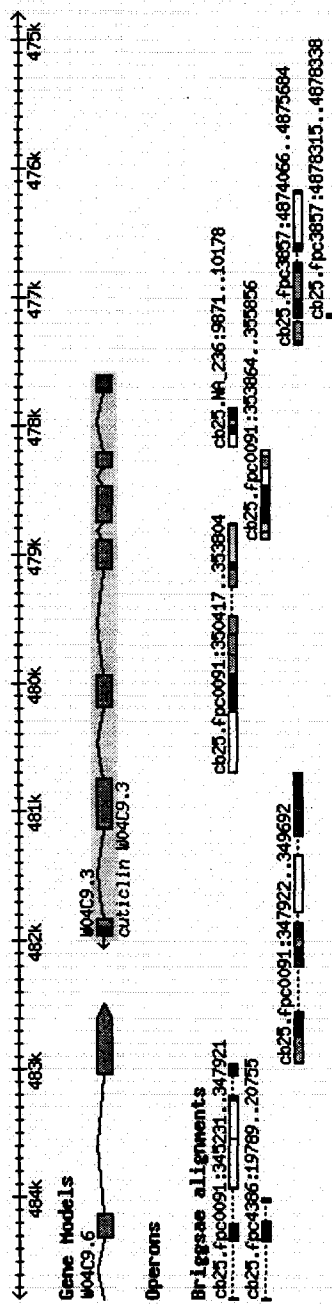
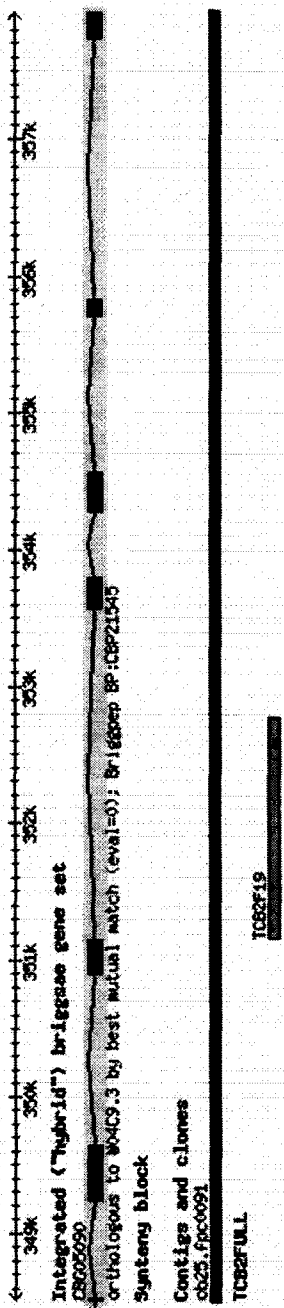
1562 **GTTCAGACAAAGTATTCAGACGCGAACAATCGGACGCGGACCGGAAATGACAGGTAAGI** tcb2_m6430
 3119 **GTTCAGACAAAGTATTCAGACGCGAACAATCGGACGCGGACCGGAAATGACAGGTAAGI** CBG10725
 GTTAC-CTC-AAAAAAGTATTCGGAATA-GTTTTTTCG-CAAGTACTGctatcttgatgttc consensus

1606 **ATGATGCGGCTTCGACGTGGAAATGCGTGGCAATATTTTCATAGTCGAGGAAAGGAAAGGAAI** tcb2_m6430
 3179 **atgatgacctttccagtgcaaatcactcggaaatatttgatgagtragcaactgaaacaat** CBG10725
 consensus

1501 **TTTTCAGACAAAGTATTCAGACGCGAACAATCGGACGCGGACCGGAAATGACAGGTAAGI** CBG10725
 1175 **TTTTCAGACAAAGTATTCAGACGCGAACAATCGGACGCGGACCGGAAATGACAGGTAAGI** CBG10725_1
 TTTTCAGACAAAGTATTCAGACGCGAACAATCGGACGCGGACCGGAAATGACAGGTAAGI consensus

3181 **GTTCAGACAAAGTATTCAGACGCGAACAATCGGACGCGGACCGGAAATGACAGGTAAGI** CBG10725
 1230 **GTTCAGACAAAGTATTCAGACGCGAACAATCGGACGCGGACCGGAAATGACAGGTAAGI** CBG10725_1
 gatcggctttccagtcgcaaatgactgcaaatatttttcgatgctcagcaacgcaaaagaaatgac consensus

no e
 ortholog
 (or blastp)
 - element
 inserts
 exactly at
 beginning
 of intron,
 ends
 before end
 of intron



```

CBG05090      267 LDEN                               CSVSRDFPQVVYLPSLTTA
                LDEN                               CSVSRDFPQVVYLPSLTTA
                LDEN                               CSVSRDFPQVVYLPSLTTA
CBG05090      4344 tggagggtgagta Intron 4 CAGAttgtcgtccggtcctcaag
                taaa <2-----[4358 : 6774]-2> gctcgatcattatcctccc
                gtgt                               ccgcaccacctcagcat
    
```

```

                270      280      290      300      310
241 DTSGYTGILLEGGLDSEGVESVPLIDENGC SVSRDFPQVVYLPSLT AYMA EAI
187 DTSGYTGILLEGGLDSEGVESVPLIDENGC SVSRDFPQVVYLPSLT AYMA EAI
                DTSGYTGILLEGGLDSEGVES-PLIDENGC SVSRDFPQVVYLPSLT AYMAIEAI-FP
    
```

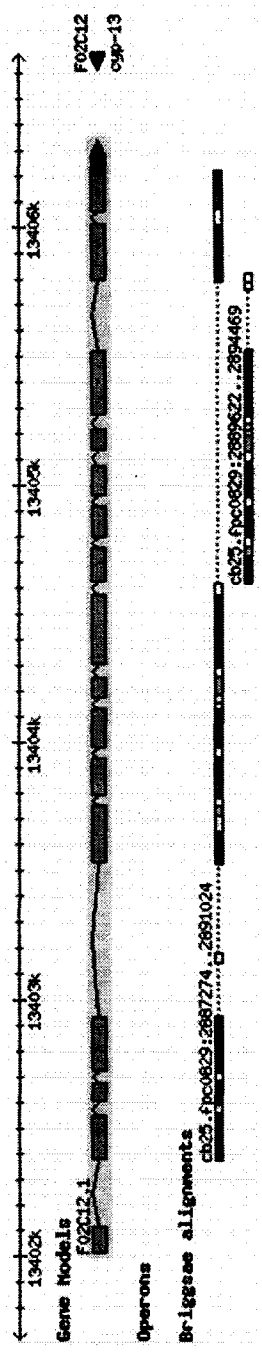
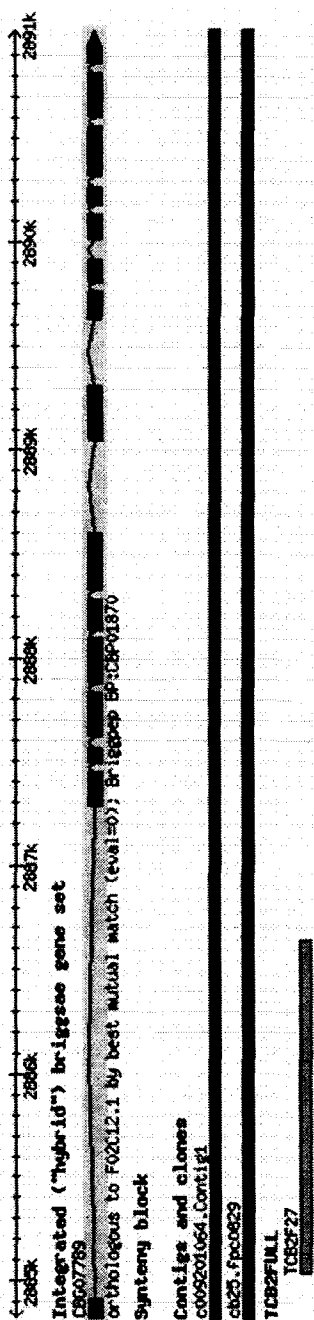
```

                2110      2120      2130      2140      2150      2160
1490 TCAGAAATGCTAGTACGACG.....CAGAAAGAAATTAAG
733 TCAGAAATG.....
2099 TCAGAAATGCTAGTACGACGTTTTTTCCATATCTATATTC
804 TCAGAAATG.....
                TCAGAAATGgtgag-a-agaggcat-----caattatra-a-ga-aatgt-c-
    
```

```

                2280      2290      2300      2310      2320
1803 TAAATAGCTCATCTCAAGGTTTACCTATCTAGAC....TTAGTCTCACT
742 .....
2219 GTTATGCAATAAGCTTCAATACGTTTCCGACACTCTTTTCCAGATCC
813 .....
                -act-t-a-a-c-g-tgg-tact-tt-g-agt-tat-t-c-atgc
    
```

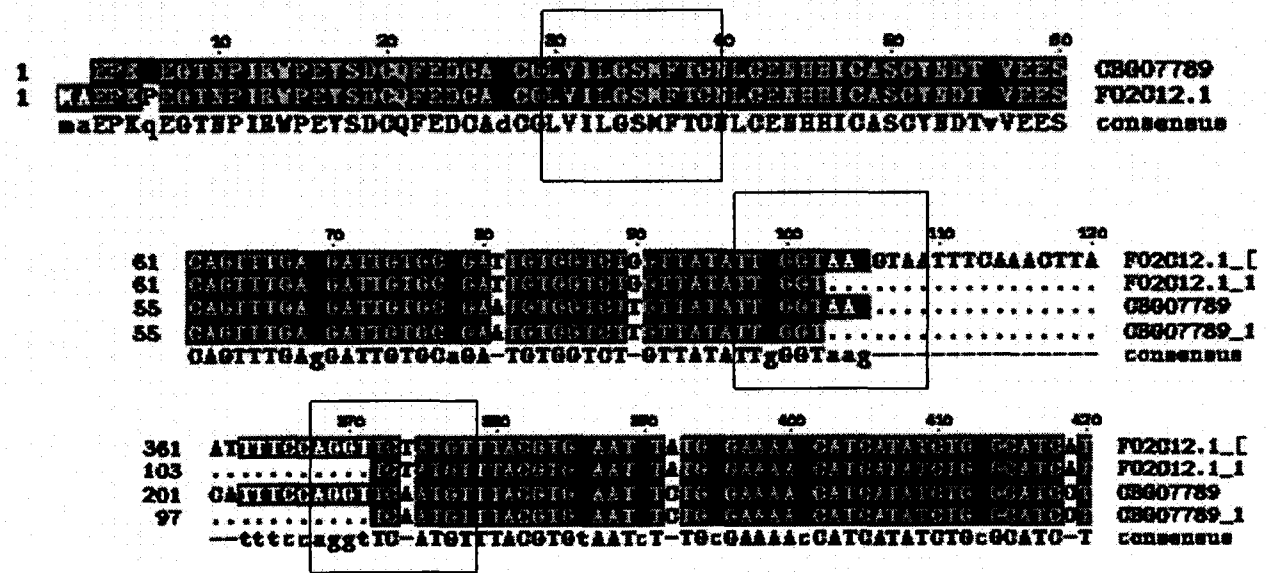
b and e end start intron at same place, but b intron ends before e (if annotation/conceptual translation) is correct

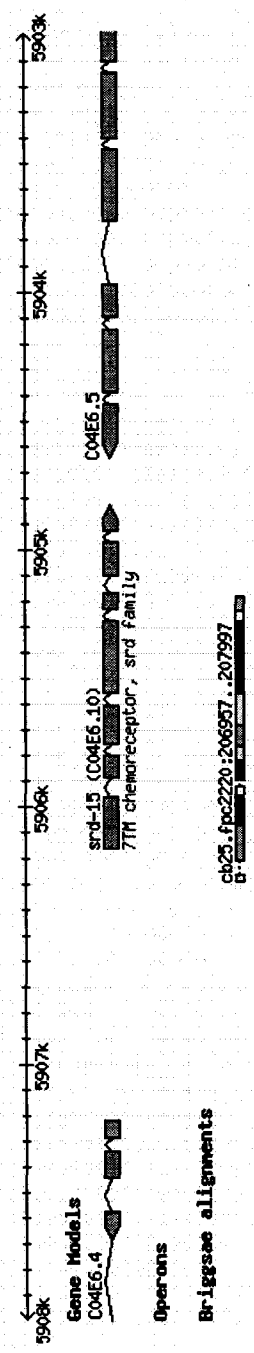
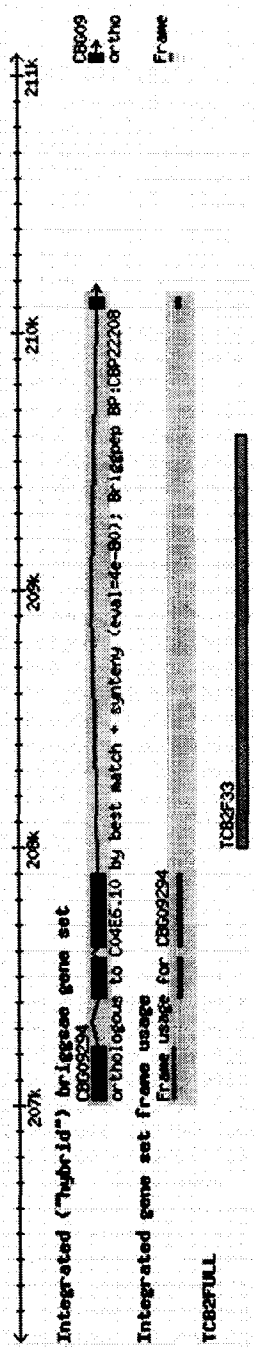


CBG07789	1	EPKQEGTNP IRWPEYSDCQFEDCAECGLVIL		
		EPKQEGTNP IRWPEYSDCQFEDCAECGLVIL		
		EPKQEGTNP IRWPEYSDCQFEDCAECGLVIL		
CBG07789	1	gcacggaacactcgttgctggtggtgcat		
		aaaagcactggcaacagataagcaggttt		
		gaaatatacagggcgctgtgtaattttag		
CBG07789	32		SMFTCNLCENHHICASCYNDTIV	
			SMFTCNLCENHHICASCYNDTIV	
		G:G[ggt]	SMFTCNLCENHHICASCYNDTIV	
CBG07789	94	GGTAAGTT Intron 1	CAGGTTatatactgacccatgttagaag	
		<1-----[95 : 2459]-1>	cttcgatgaaaatgccgaaactt	
			agtgttccacttccaccctcatt	

b and e conserved
intron, b larger due to
element

175





b protein from wise ends after
 exon 2 – prediction doesn't
 include last several aa, but can
 see same start of intron as below

CEG09294 195 SRNVMSKETKAVHAQLLK
 SRNVMSKETKAVHAQLLK
 SRNVMSKETKAVHAQLLK
 CEG09294 818 acagatagaaggcgctca
 ggattcaacactacatta
 tatggagaaattttaaga

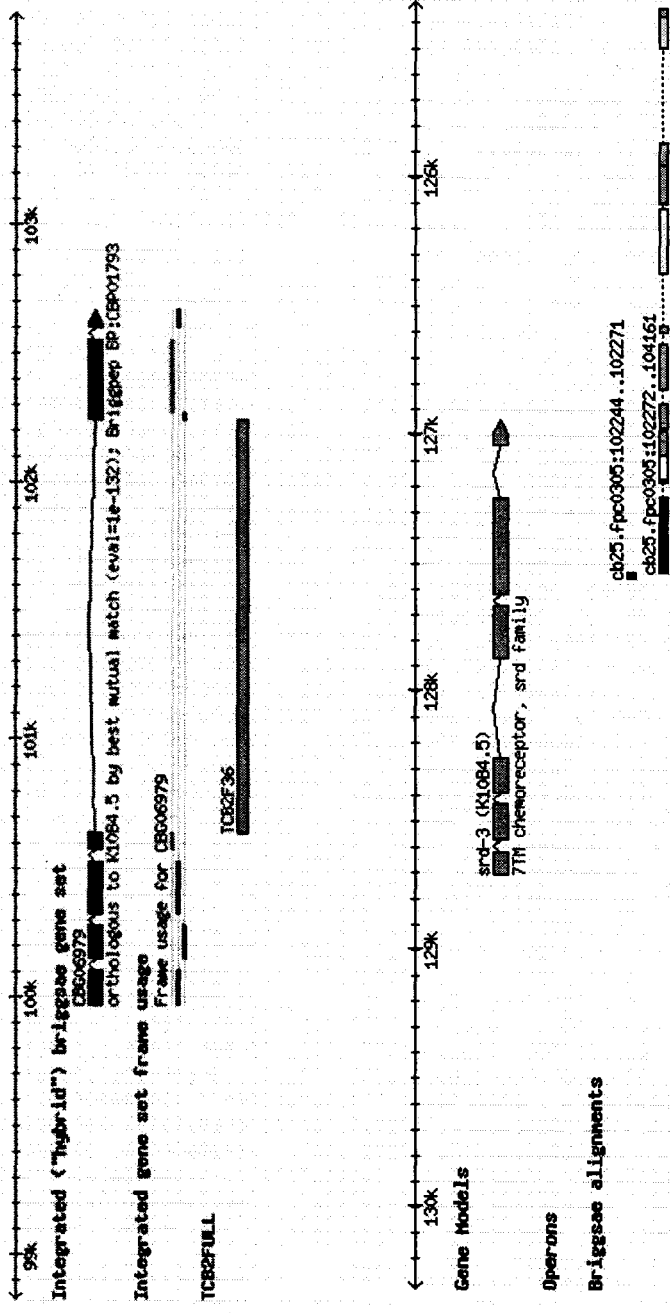
152 1GTHLESVSA YA LEMT EYIPVI IPIERZI IIPSPYKSEKHA EAQL
 161 1GTHLESVSA YA LEMT EYIPVI IPIERZI IIPSPYKSEKHA EAQL
 170 1TGHILFVSAIYVLENTIPYIPYIIGIFILRNLIK-LK-S-IVASKEIHAHQLS
 210 CEG09294
 00426.10
 consensus

212 **SRNVMSKETKAVHAQLLK**
 241 **SRNVMSKETKAVHAQLLK**
 I-I-fqafipvaav--vyyig-qigigizayavstvfipllspvayivpyr
 CEG09294
 00426.10
 consensus

b and e similar, but
 intron not conserved
 – starts at the same
 point, but e exon
 starts earlier

829 **GAITAGT** GAAIGI AIGIGIA GAAAGAA GCA TIGIIGGICA IT
 855 **GAITAGT** GAAIGI AIGIGIA GAAAGAA GCA TIGIIGGICA IT
 814 **GAGAGT** MAIIGI AIGIGIA GAAAGAA GGT TIGIIGGICA IT
 579 **GAGAGT** MAIIGI AIGIGIA GAAAGAA GGT TIGIIGGICA IT
 6A-GAGTGG-AATGTGATGTC-AAGAAAGAAAAGC-GTTCATGCTGAAATTA-TGAAAGT
 900 CEG09294
 00426.10_L
 00426.10_1
 CEG09294
 CEG09294_1
 consensus

1063 **GAGTGC** AATIGGIA AATGATIGGIA AATGATIGGIA AATGATIGGIA
 788 ...UGCAATIGGIA AATGATIGGIA AATGATIGGIA AATGATIGGIA
 1011 ACTA **GAAT** GATGATIGGIA AATGATIGGIA AATGATIGGIA
 637AATGATIGGIA AATGATIGGIA AATGATIGGIA
 ---tgcattt85-1c5-10-GTGGAAI-gI-I-ct85-gtgr80I---ctctacagtaa
 1100 CEG09294
 00426.10_L
 00426.10_1
 CEG09294
 CEG09294_1
 consensus



CBG06979 170 SPSTFAAIMY --MTIPCFPYAVI
 SPSTFAAI+Y MTIPCF YAVI
 SPSTFAAIY LYMTIPCF!YAVI
 CBG06979 645 tctatggaatAGTACTGG Intron 4 CAGTactaaactt4tggg
 ccctcctta <1-----[676 : 2273]-1> tatctcgt actt
 accattgacc gcgggttcc cagt

152 H I K E A D Y I A G Y S S E P S P P A I V I T I P C F Y V I V I E R A A I I G G G R T I P S CBG06979
 181 S E E A G V I A G Y S Q S A A E S A I E N H G S P I Y V I V F S E R T I E S D G G R I I E S K1084.5
 d-KE1VGDLYIAGYMS1VSR-IV-AIVYTIPOPPHYGVIV-FRYK-LK-LDQIGR-INS consensus

1134 T T I G C A T C T C K1084.5_[1]
 606 T T I G C A T C T C K1084.5_[6]
 658 T T I G C A T C T C A G T A C T G C G A A A A A G G T A T T C G C A A A C G A T T T T T G A G A C T A CBG06979
 521 T T I G C A T C T C CBG06979_1
 Tc-CTGGAT-ATC consensus

e has no
 intron,
 element
 forms
 intron,
 see next
 slide

1145 K1084.5_[1]
 622 K1084.5_[6]
 713 A C T T T G C T G A A A C T T A G C G A T C G A T C G A A A A A A A A A T A T C A G A T T G A T C C G G G A A G T T T C A CBG06979
 534 CBG06979_1
 consensus

1145 K1084.5_[1]
 622 G A A A G T T A T C G C G A A A A A C T T A T T C G G A A T A C T T T T T G C C A T C G A T C G A A A A A A A A A T A T C A G A T T G A T C C G G G A A G T T T C A K1084.5_[6]
 534 CBG06979
 G T A C A T G A T C consensus

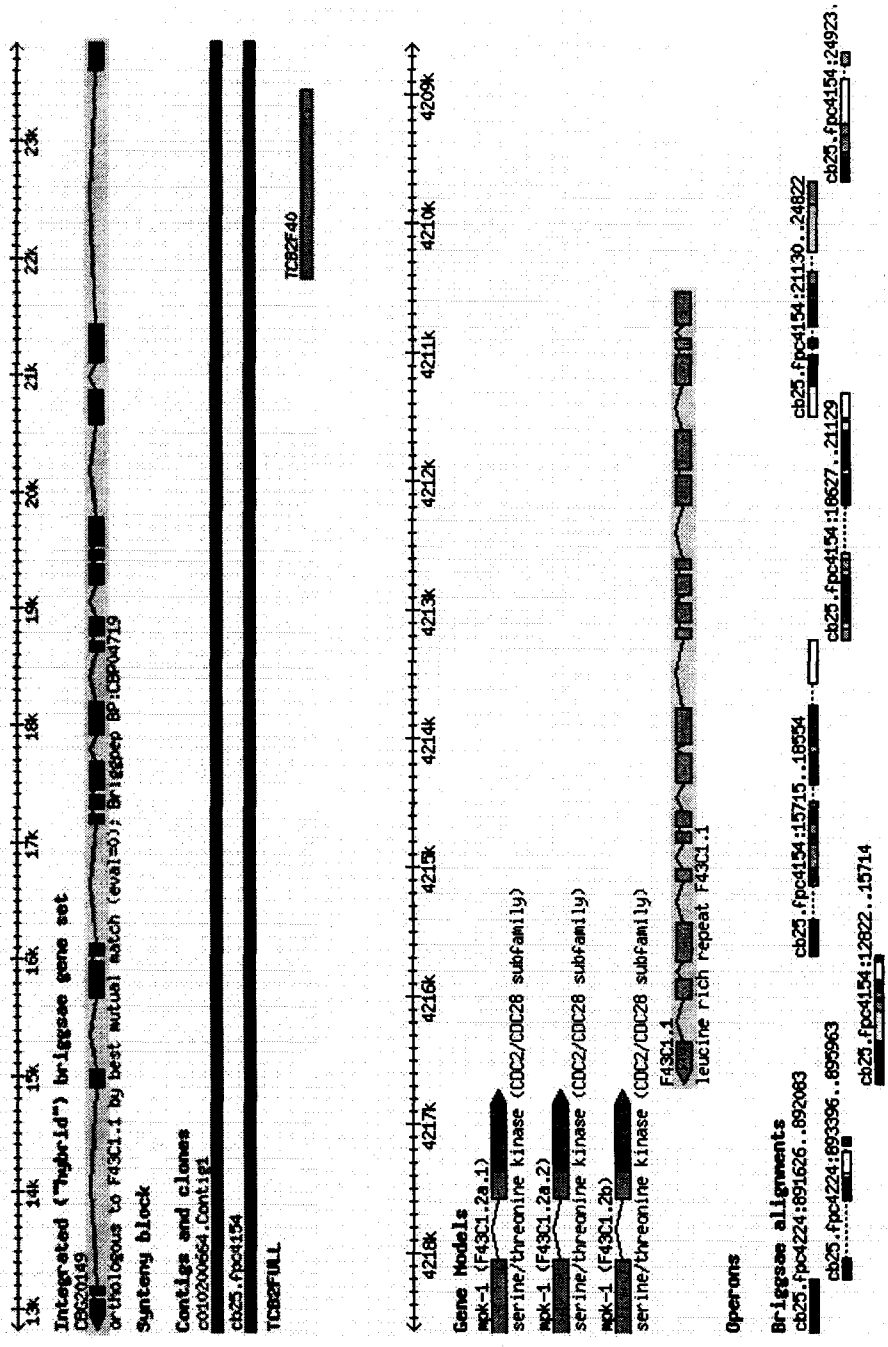
660 TGG ATATAGTCTAATGCTAMIGGCGAAMAGGTAATAGCAACG TTTTITTTGAGT . TTTAA
 1 TACTATGIGGGGAAAAGGTAATAGCAACG TTTTITTTGAGT . TTTAA
 64 TGG TATATGGT TGG . . . TTTTITTTGAGT . TTTAA
 attg-a---atc---GAGT acg ggcraaaaAGGTATTGcraaa cgg TTTTttGaga-GTAA

cbg06979_c
 tcb2_m6430
 WBGeme0000
 consensus

2236 GGTGGAAATGDTATT GGGATA GTTTTGGGGGAGTGTGAC TGTGATTT . . . TTTG
 1567 TCCTGAAAACIATTTGCGAATAGCCTTTTGGCGAGTACIG
 1606 TTTCTTTAAA . . . TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT
 gcTgaaAAAcTattTtcgAAATatcttttTggcGagTactTgt---gtg-aatt---tttt-

cbg06979_c
 tcb2_m6430
 WBGeme0000
 consensus

intron is element!



b and e conserved
intron, b larger due to
element

```

CBG20149      50  SEEVIRKGLIHSWQNAARAVFYNIRLQT
                SEEVIRKGLIHSWQNAARAVFYNIRLQT
                SEEVIRKGLIHSWQNAARAVFYNIRLQT
CBG20149      148  aggggaaagacacacagaggttataaacca
                gaattgagattaggaacgcttaacatgtac
                taaggagggtctgatgacgttcgtcggat
CBG20149      81  NNQDEALQIRNMKVVDIYESKKGR
                NNQDEALQIRNMKVVDIYESKKGR
                NNQDEALQIRNMKVVDIYESKKGR
CBG20149      241  AGTAAGTT Intron 1  CAGAAaacgggccacaaggatgtaagc
                <1-----[242 : 2407]-1>  aaaaactatgatataaacaagg
                ctacattacgtgggcttgaaaagg

```

```

59  KLIEH VQARAVFYNIRLQIKKIKKQVVALQIEEKKYD ESK QRSRLKDDDF LII  CBG20149
61  KLIEH VQARAVFYNIRLQIKKIKKQVVALQIEEKKYD ESK QRSRLKDDDF LII  F43G1.1
        KLIEH VQARAVFYNIRLQIKKIKKQVVALQIEEKKYD ESK QRSRLKDDDF LII  consensus

```

```

235  CAAGC AGT AGTITTTTCTGCTAGAG AAT CA A AAT ATTCGGA AAT  CBG20149
235  CAAGC A.....  F43G1.1
241  CAAGC AGT AGTITTTTCTGCTAGAG AAT CA A AAT ATTCGGA AAT  CBG20149_1
241  CAAGC A.....  F43G1.1
        CAAGCagtgagtt-ttg 8---c-ga-caatcag-gagaa-c8888---t---aa-- consensus

```

```

355  CAGAAAGCA CAGCA GA GCCTTCAGAT CC AATATAA GT GA TA ATGAGI  CBG20149
242  ...AAGACA CAGCA GA GCCTTCAGAT CC AATATAA GT GA TA ATGAGI  CBG20149_1
361  CAGAAAGCA CAGCA GA GCCTTCAGAT CC AATATAA GT GA TA ATGAGI  F43G1.1
243  ...AAGACA CAGCA GA GCCTTCAGAT CC AATATAA GT GA TA ATGAGI  F43G1.1
        CAGAAAGCA CAGAAAGCAATTCAGAAATCAGAAATATAAAGTAAGTGT-CATGAGT consensus

```