



## **Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin**

### **Copyright statement**

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

### **Liability statement**

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

### **Access Agreement**

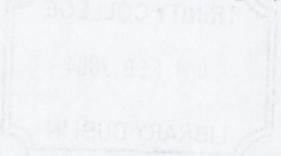
By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Evolution of the genomes of two  
nematodes in the genus  
*Caenorhabditis*

by  
Avril Coghlan

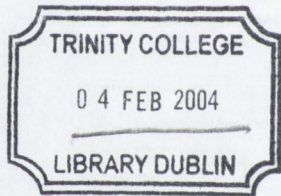
A thesis submitted to  
The University of Dublin  
for the degree of  
Doctor of Philosophy



Supervised by Professor Kenneth H. Wolfe  
Department of Genetics  
Trinity College  
University of Dublin

October, 2003





THESIS  
7844✓

# Declaration

This thesis is submitted by the undersigned for the degree of Doctor of Philosophy at the University of Dublin. It has not been submitted as an exercise for a degree in any other university.

Apart from the advice, assistance, and joint effort mentioned in the acknowledgements and in the text, this thesis is entirely my own work.

I agree that the library may lend or copy this thesis freely upon request.

Avril Coghlan. October, 2003.

*avril coghlan*



# Acknowledgements

Thankyou a billion Mum, Dad and Noel for always cheering me on during my Ph.D..

I am hugely grateful to Ken — thank you so much for your inspiration and encouragement.

Thankyou Andrew for being willing to read manuscripts and discuss ideas.

A special thanks to Karsten, Kevin and Simon, who kept my computer alive.

Thankyou to Richard Durbin, who allowed me to join the *C. briggsae* Sequencing Project, and to the people with whom I worked on the project, especially Marc Sohrmann, Todd Harris, and Lincoln Stein.

Thanks to Dónall MacDónaill and Nigel Buttimore, for your enthusiasm in our collaboration at the start of my Ph.D..

Thanks to all the really great members of the Wolfe lab. — it wouldn't have been a quarter as much fun without you: Aoife, Karsten, Simon, Cathal, Antoinette, Kevin, Guillaume, Mario, Brian, Lucasz, Greg, Ailís, Devin, and Jonathen.

Many thanks to my thesis examiners, Mark Blaxter and Dan Bradley, for detailed and insightful discussion of my results.

# Summary

The soil-dwelling nematode *Caenorhabditis elegans* has been intensively studied as a model organism for the last 40 years. It was the first animal for which we had a complete description of development, anatomy, a neural wiring diagram, and, in 1998, a genome sequence. In 2001 the genome of *Caenorhabditis briggsae* was sequenced. They are the first pair of animals from the same genus to have their genomes sequenced. The two worms are very similar morphologically and follow similar developmental programs, but are surprisingly dissimilar genetically. I compared their genomes to identify syntenic regions that have been conserved since they diverged 80–110 million years ago. I found the rate of chromosomal rearrangement to be exceptionally high in these nematodes compared to in most eukaryotes. After the *C. briggsae* genome was sequenced in 2001, an important step was the prediction of protein coding genes in the raw sequence. I describe how my collaborators and I predicted genes in the *C. briggsae* genome; compared *C. briggsae* genes to those of *C. elegans*; and used similarity to *C. briggsae* to improve gene predictions in *C. elegans*. Intron-exon structure has evolved rapidly: I estimated there have been 0.005 intron gains or losses per gene per million years since the two species diverged. To elucidate the mechanism of intron gain, I pinpointed intron-exon differences that were due to intron gain rather than loss. My results narrow down the probable mechanism of intron gain to just two of the five hypothesised mechanisms.



# Table of Contents

List of Abbreviations .....	vii
1. Overview .....	1
1.1 The Phylum Nematoda .....	1
1.2 The Origin of Nematodes .....	2
1.3 A Model Nematode, <i>Caenorhabditis elegans</i> .....	2
1.4 Two <i>Caenorhabditis</i> Genomes .....	5
1.5 The <i>Brugia malayi</i> Genome .....	5
1.6 Nematode Genomes: the Future? .....	5
2. Faster Rate of Genome Rearrangement in Nematodes than <i>Drosophila</i> ..	7
2.1 Introduction .....	7
2.2 Results .....	9
2.2.1 Detection of Conserved Segments and their Length Distribution .....	9
2.2.2 Differences among and along <i>C. elegans</i> Chromosomes .....	11
2.2.3 Estimating the <i>C. briggsae</i> - <i>C. elegans</i> Divergence Date .....	11
2.2.4 Duplications .....	12
2.2.5 Rates of Reciprocal Translocation, Inversion, and Transposition .....	12
2.2.6 Association of Breakpoints with Repetitive DNA .....	17
2.3 Discussion .....	19
2.4 Future Work .....	23
2.5 Methods .....	24
2.5.1 Sources of Sequence Data .....	24
2.5.2 Predicting <i>C. briggsae</i> Genes .....	25
2.5.3 Finding Orthologues .....	25
2.5.4 Estimating the <i>C. briggsae</i> - <i>C. elegans</i> Divergence Date .....	26

2.5.5 Finding Conserved Segments and Classifying Breakpoints .....	27
2.5.6 Testing Whether Breakpoints are Associated with Repeats .....	28
2.6 Acknowledgements .....	28
3. The <i>Caenorhabditis briggsae</i> Gene Set .....	29
3.1 Introduction .....	29
3.2 Results .....	30
3.2.1 Protein Coding Genes .....	30
3.2.2 Comparing the <i>C. briggsae</i> and <i>C. elegans</i> Gene Sets .....	32
3.2.3 <i>C. briggsae</i> - <i>C. elegans</i> Orthologues .....	33
3.2.4 Estimating the <i>C. briggsae</i> - <i>C. elegans</i> Divergence Date .....	34
3.2.5 <i>C. briggsae</i> - <i>C. elegans</i> Paralogues and Orphans .....	34
3.2.6 Using <i>C. briggsae</i> Sequence to Improve <i>C. elegans</i> Annotation .....	35
3.3 Discussion .....	36
3.4 Future Work .....	38
3.5 Methods .....	38
3.5.1 Protein coding Gene Prediction .....	38
3.5.2 Finding <i>C. briggsae</i> - <i>C. elegans</i> Orthologues .....	40
3.5.3 Detecting Intron Gain and Loss in Orthologues .....	40
3.5.4 Estimating the <i>C. briggsae</i> - <i>C. elegans</i> Divergence Date .....	41
3.5.5 Using <i>C. briggsae</i> Sequence to Improve <i>C. elegans</i> Annotation .....	42
3.6 Acknowledgements .....	42
4. Origins of Novel Introns in <i>Caenorhabditis</i> .....	43
4.1 Introduction .....	43
4.2 Results .....	44
4.2.1 Identification of Novel Introns .....	44
4.2.2 Exon Splice Site Consensus of Novel Introns .....	45



4.2.3 Phases of Novel Introns .....	46
4.2.4 Germline Expression of Genes that have Gained Introns .....	46
4.2.5 Repeat Elements in Novel Introns .....	47
4.3 Discussion .....	47
4.3.1 Method for Identifying Recently Gained Introns .....	47
4.3.2 Rate of Intron Gain in <i>C. elegans</i> vs. <i>C. briggsae</i> .....	49
4.3.3 Mechanisms of Intron Gain .....	49
4.3.4 Germline Expression .....	50
4.3.5 The Proto-Splice Site .....	51
4.3.6 The Molecular Smoking Gun .....	52
4.3.7 Conclusion .....	52
4.4 Methods .....	53
4.4.1 Sources of Sequence Data .....	53
4.4.2 Finding the Closest Homologues of each Nematode Gene .....	53
4.4.3 Detecting Intron Gains from Protein Alignments .....	53
4.4.4 Checking whether Putative Novel Introns are Present in <i>Brugia malayi</i> .....	54
4.4.5 Phylogenetic Support for Intron Gains .....	54
4.4.6 Control Set of Introns .....	55
4.4.7 Detecting Repeat Elements in Introns .....	55
4.5 Acknowledgements .....	56
List of References .....	57

# Abbreviations

**BLAST** basic local alignment search tool

**BLOSUM** blocks substitution matrix

**bp** basepairs

**cDNA** copy DNA

**DNA** deoxy-ribonucleic acid

**EST** expressed sequence tag

**HSP** high-scoring segment pair

**kb** kilobases

**Mb** megabases

**mRNA** messenger RNA

**Mya** million years ago

**Myr** million years

**RNA** ribonucleic acid

**rRNA** ribosomal RNA

**ORF** open reading frame

**OST** ORF sequence tag

**TIGR** The Institute for Genomic Research

**TIR** terminal inverted repeat



# List of Figures

1.1 The model organism <i>Caenorhabditis elegans</i> .....	2
1.2 The relationships of the animal phyla .....	3
1.3 A phylogenetic tree of the phylum Nematoda .....	4
2.1 Distribution of sizes of conserved segments .....	10
2.2 The <i>Caenorhabditis elegans</i> region surrounding the <i>sex-1</i> locus .....	10
2.3 Location of conserved segments in the <i>Caenorhabditis elegans</i> genome .....	13
2.4 Estimates of the <i>C. briggsae</i> - <i>C. elegans</i> speciation date .....	14
2.5 Sizes of inversions and transpositions .....	15
2.6 Method of detecting inversions and transpositions .....	16
3.1 Joint refinement of <i>C. elegans</i> and <i>C. briggsae</i> gene models .....	31
3.2 A region on <i>C. elegans</i> chromosome III, and the syntenic <i>C. briggsae</i> region .....	37
4.1 Identifying novel introns .....	45
4.2 The exon splice site consensus of novel introns .....	46
4.3 An example of a novel intron containing a repeat element .....	48

# List of Tables

2.1 Association of rearrangement breakpoints with repeats .....	17
2.2 Association of translocation and transposition breakpoints with particular repeats .	18
3.1 Comparison of the <i>C. briggsae</i> and <i>C. elegans</i> gene sets .....	33
3.2 Updating the <i>C. elegans</i> gene set using <i>C. briggsae</i> similarity .....	35

---

---



# Chapter 1

## Overview

The biology of the model nematode *Caenorhabditis elegans* is reviewed in detail in the book *C. elegans II* (Riddle et al., 1997). In this chapter I briefly introduce the Phylum Nematoda and nematode genomes. Each of the three topics researched for my Ph.D. is introduced in detail at the start of Chapters 2, 3 and 4.

### 1.1 THE PHYLUM NEMATODA

*If all the matter in the universe except the nematodes were swept away, our world would still be dimly recognisable, and if, as disembodied spirits, we could then investigate it, we should find its mountains, hills, vales, rivers, lakes, and oceans represented by a film of nematodes. . . (Cobb, 1915)*

Nematodes are non-segmented invertebrates that have a body cavity, a digestive tract, a nervous system, an excretory system, and a set of longitudinal muscles, but lack any appendages. Most nematodes are microscopic; the model organism *Caenorhabditis elegans* is just 1 mm long (Figure 1.1). In terms of the numbers of individuals, nematodes are the most abundant type of animal on earth (Chitwood and Chitwood, 1974; Andr assy and Zombori, 1976). So far 20,000 species have been classified, and there may be up to ten million species (Blaxter, 1998). This abundance results from their ability to adapt, and is due to a small size, a resistant cuticle, and a simple body plan. Small changes to their body plan have allowed invasion of many different habitats. Nematodes live in hot springs, polar ice, soil, fresh and salt water, and as parasites of plants, insects, vertebrates, and other nematodes (Andr assy and Zombori, 1976). This evolutionary plasticity has long fascinated biologists. However, a more urgent reason to study them is the damage they cause to human health and agriculture. Over 3.5 billion people are infected by nematodes, while each year plant parasitic nematodes cause about \$100 billion of damage to crops (Lilley et al., 1999; Luong, 2003).





Figure 1.1: The model organism *Caenorhabditis elegans*: an adult with two juveniles.

## 1.2 THE ORIGIN OF NEMATODES

The first nematode was probably a free-living marine animal that fed on bacteria (Poinar, 1983). The relationship of this early nematode to other animals is uncertain. Based on phylogenetic trees, some argue they are most closely related to arthropods (Aguinaldo et al., 1997; Mushegian et al., 1998; Figure 1.2). Conflicting phylogenetic results suggest they are an outgroup to a clade that includes arthropods and chordates (Blair et al., 2002).

Doubts also surround when the first nematode lived. Related animal phyla such as the arthropods appear in the fossil record ~550 million years ago (Mya). However, because they lack hard body structures, nematode fossils are scarce, and most fossils found are recent, from 20–120 Mya (Poinar, 1983). Based on both fossil and phylogenetic evidence, paleontologists believe that the animal kingdom split into a few basal clades ~1000 Mya, and that later, ~600–800 Mya, animal groups acquired new body plans, giving rise to the modern phyla including the nematodes and arthropods (Benton and Ayala, 2003).

## 1.3 A MODEL NEMATODE, *Caenorhabditis elegans*

*Caenorhabditis elegans* belongs to the order Rhabditida: small (1–2 mm), free-living worms that feed on bacteria in decaying organic matter. The closest parasitic relatives of the rhabditids are the strongylids, small gut parasites such as the human hookworm *Necator americanus* (Blaxter et al., 1998; Figure 1.3).

In 1963 Sydney Brenner realised that species from the genus *Caenorhabditis* would be ideal model organisms. They have many cell types involved in complex functions in mammals, including intestine, muscle and excretory cells, and neurons. Since then study of *C. elegans* has contributed to understanding of central biological processes: apoptosis, signalling pathways, cell movement and polarity, sex determination, and synaptic signalling (Riddle et al., 1997). *C. elegans* was the first animal for which we had a complete description of development, anatomy, and a neural wiring diagram; and, as a crowning glory, the first to have its genome sequenced (Riddle et al., 1997; The *C. elegans* Sequencing Consortium, 1998).



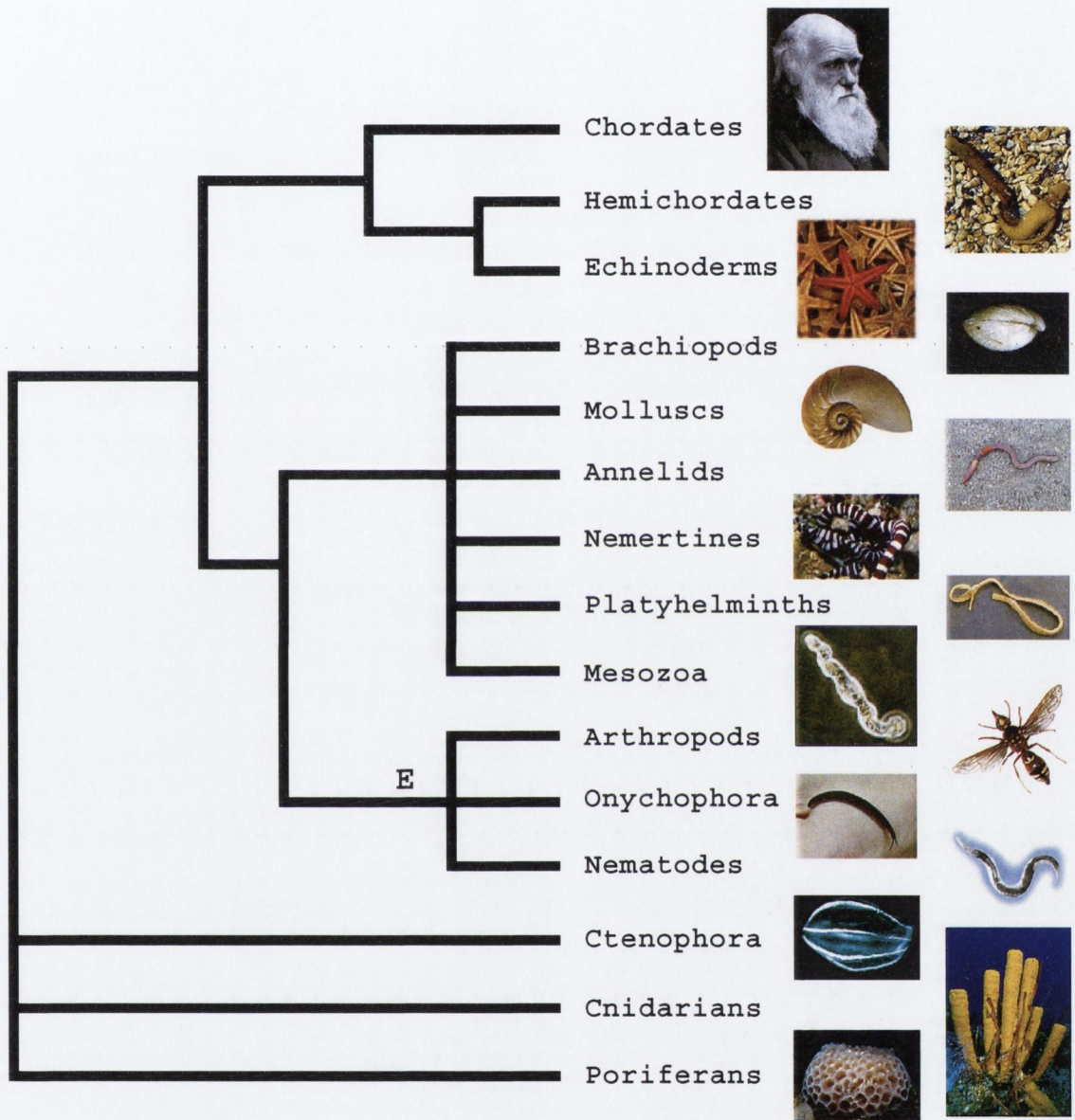


Figure 1.2: The relationships of the animal phyla, according to Aguinaldo et al. (1997), who hypothesised that nematodes and arthropods belong to a clade of moulting animals, the Ecdysozoa (shown as *E*).

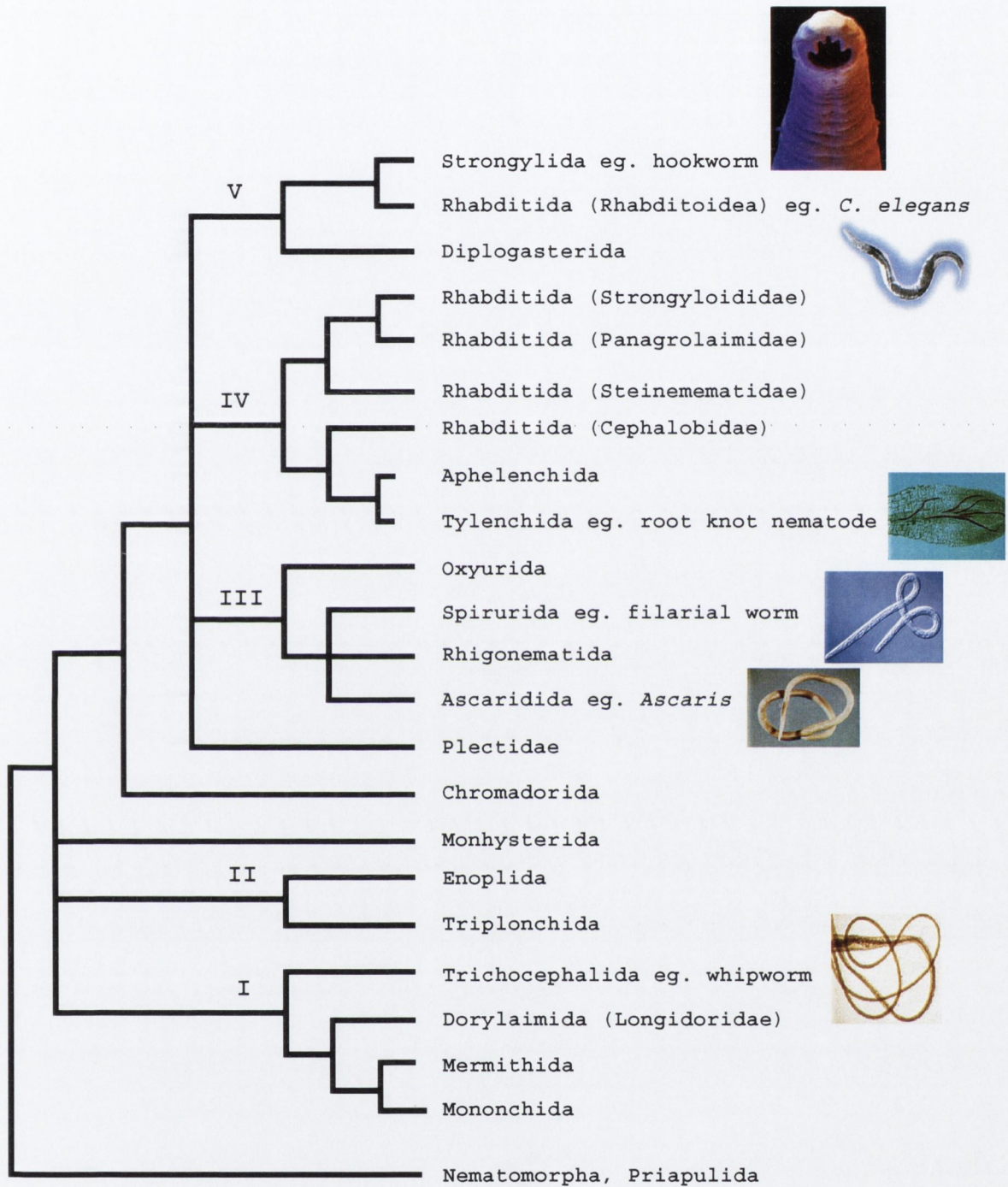


Figure 1.3: A phylogenetic tree of the phylum Nematoda, from Blaxter et al. (1998). The model nematode *C. elegans* belongs to clade V. The filarial nematode, *Brugia malayi*, whose genome is currently being sequenced, belongs to the order Spirurida in clade III. Clades V and III diverged about 550 Mya (Vanfleteren et al., 1994).



## 1.4 TWO *Caenorhabditis* GENOMES

In 2001 the genome of *C. briggsae* was sequenced (Stein et al., 2003). *C. elegans* and *C. briggsae* are the first two animals from the same genus to have their genomes sequenced. As part of my Ph.D. work, I was fortunate to collaborate on the *C. briggsae* Sequencing Project (Stein et al., 2003). The work I did, predicting the *C. briggsae* gene set, is described in Chapter 3.

*C. elegans* and *C. briggsae* are very similar morphologically and follow similar developmental programs in, for example, sex determination and vulval development (Nigon and Dougherty, 1949; Stothard and Pilgrim, 2003; Kirouac and Sternberg, 2003). However, although both have six chromosomes and similar genome sizes (~100–105 Mb), they are surprisingly dissimilar genetically (Stein et al., 2003). They diverged 80–110 Mya, approximately when primates split from rodents. However, they have diverged more rapidly than human and mouse have, by both chromosomal rearrangements and nucleotide substitutions (Stein et al., 2003). In Chapter 2, I describe how we compared the *C. elegans* and *C. briggsae* genomes, and discovered that their rate of chromosomal rearrangement is exceptionally high compared to that of most eukaryotes (Coghlan and Wolfe, 2002).

## 1.5 THE *Brugia malayi* GENOME

The *C. elegans* and *C. briggsae* genomes will soon be joined by that of a third nematode, *Brugia malayi*, a parasite which causes lymphatic filariasis. The 110 Mb genome of *B. malayi* is being sequenced by The Institute for Genomic Research (TIGR; <http://www.tigr.org/tdb/e2k1/bma1/>). Already the sequence is > 95% complete (E. Ghedin, pers. comm., March 2003). *B. malayi* is a distant relative of *C. elegans* and *C. briggsae* (Figure 1.3). Thus, to molecular evolutionists the *B. malayi* genome will be the perfect outgroup for distinguishing whether *C. briggsae*-*C. elegans* differences are due to *C. elegans*- or *C. briggsae*-specific changes.

## 1.6 NEMATODE GENOMES: THE FUTURE?

In 1965, the German zoologist Alfred Kaestner wrote that “our knowledge concerning the evolution of nematodes is next to nothing.” Happily, with three nematode genome sequences in hand, now our knowledge is growing very fast. However, perhaps we have more questions than answers: we do not know why the arms of *C. elegans* chromosomes evolve faster than the centres; why most clade V nematodes have six chromosomes despite numerous genome rearrangements; nor what is the function of the ~1000 *C. briggsae* genes that lack any *C. elegans* sequence match (Stein et al., 2003)? The data set of three genomes is ideal for tackling questions about nematode evolution. But they can also be studied to address questions relevant to all eukaryotes. An example is our investigation of how new introns arise, by comparing the *C. elegans*, *C. briggsae* and *B. malayi* genomes, described in Chapter 4. Looking forward,

it seems very possible that once again these tiny animals will be first in revealing some of nature's deepest secrets.



## Chapter 2

# Faster Rate of Genome Rearrangement in Nematodes than *Drosophila*

The research described in this chapter was published in *Genome Res.* (Coghlan and Wolfe, 2002). Our results were based on the 13% of the *C. briggsae* genome that had been sequenced by 2001. In this chapter I extend the discussion of the paper, to comment whether our results agree with later analysis of the whole *C. briggsae* genome (Stein et al., 2003). I have also added a section on future work.

### ABSTRACT

We compared the genome of the nematode *Caenorhabditis elegans* to 13% of that of *C. briggsae*, identifying 252 conserved segments along their chromosomes. We detected 517 chromosomal rearrangements, with the ratio of translocations to inversions to transpositions being  $\sim 1:1:2$ . We estimate that the species diverged 50–120 million years ago, and that since then there have been 4030 rearrangements between their whole genomes. Our estimate of the rearrangement rate, 0.4–1.0 chromosomal breakages/Mb per Myr, is at least four times that of *Drosophila*, which was previously reported to be the fastest rate among eukaryotes. The breakpoints of translocations are strongly associated with dispersed repeats and gene family members in the *C. elegans* genome.

## 2.1 INTRODUCTION

The genes of *Caenorhabditis elegans* appear to have an unusually rapid rate of evolution. The substitution rates of many *C. elegans* genes are twice those of their orthologues in non-nematode metazoans (Aguinaldo et al., 1997; see Figure 3 in Mushegian et al., 1998). Even among nematodes, the *C. elegans* small subunit



ribosomal RNA gene evolves faster than its orthologues in most of the major clades (see Figure 1 in Blaxter et al., 1998). It has been estimated that two-thirds of *C. elegans* protein coding genes evolve more rapidly than their *Drosophila* orthologues (Mushegian et al., 1998). In vertebrates at least, the rate of nucleotide substitution is correlated with that of chromosomal rearrangement (Burt et al., 1999).

Ranz et al. (2001) reported that *Drosophila* chromosomes rearrange at least 175 times faster than those of other metazoans, and at a rate at least five times greater than the rate of the fastest plant genomes. However, no *Caenorhabditis* rate data existed to compare with the *Drosophila* data. Given their fast rate of nucleotide substitution, we guessed that *Caenorhabditis* genomes might have a fast rate of rearrangement. Here, we have estimated the rate of rearrangement since the divergence of *C. elegans* from *Caenorhabditis briggsae*, using the complete *C. elegans* genome sequence (The *C. elegans* Sequencing Consortium, 1998) and 13 Mb of the *C. briggsae* genome sequenced by the Washington University Genome Sequencing Center (<http://genome.wustl.edu/gsc/>). Previous studies have shown that *C. elegans* and *C. briggsae* have conserved segments of  $\geq 6$  genes (Kuwabara and Shah, 1994; Thacker et al., 1999).

To calculate the rate, we estimated the number of chromosomal rearrangements since the speciation of *C. elegans* and *C. briggsae*. Because both species have six chromosomes (Nigon and Dougherty, 1949), we assumed that there have not been any fusions or fissions of whole chromosomes since they diverged. Kecioglu and Ravi (1995), Hannenhalli (1996), and Pevzner and Tesler (2003a) have developed algorithms that deduce the historical order and sizes of the reciprocal translocations (whereby two nonhomologous chromosomes exchange chunks of DNA by recombination) and/or inversions that have occurred since the divergence of two multichromosomal genomes. However, the *C. elegans* genome evolves not only by reciprocal translocations and inversions, but also by transpositions (whereby a chunk of DNA excises from one chromosome and inserts into a nonhomologous chromosome) and duplications (Robertson, 2001). We designed a simple algorithm to calculate the number and sizes of such mutations, although not the order in which they occurred. Our method starts by finding all perfectly conserved segments between two species, in which gene content, order and orientation are conserved. Next, these segments are fused into larger segments that have been splintered by duplications, inversions, or transpositions. When no more segments can be merged, the final fused segments are assumed to have resulted from fission of chromosomes by reciprocal translocations.

To convert the observed number of rearrangements into a rate, it is necessary to have an accurate estimate of the *C. briggsae*-*C. elegans* divergence date. Emmons et al. (1979) were the first to estimate this date, using restriction fragment data, venturing that it must be "tens of millions of years" ago. Butler et al. (1981) speculated that the date was 10–100 million years ago (Mya), judging from 5S rRNA sequences, anatomical differences, and protein electrophoretic mobilities. Subsequent estimates based on sequence data were 30–60 Mya (Prasad and Baillie, 1989, one gene), 23–32 Mya (Heschl and Baillie, 1990, one gene), 54–58 Mya (Lee et al., 1992, two genes), and 40 Mya (Kennedy et al., 1993, seven genes).



Nematode fossils are extremely scarce (Poinar, 1983). Therefore, to calibrate the molecular clock, these studies either assumed that all organisms have the same silent substitution rate (Prasad and Baillie, 1989; Heschl and Baillie, 1990) or nonsilent substitution rate (Lee et al., 1992), or that *C. elegans* has the same silent rate as *Drosophila* (Kennedy et al., 1993). These are dubious assumptions; for example, Mushegian et al. (1998) showed that about two-thirds of *C. elegans* genes have a higher rate of nonsilent substitution than their orthologues in *Drosophila*. To gain a more reliable interval estimate of the *C. briggsae*-*C. elegans* speciation date, we used phylogenetic analysis of all genes for which orthologous sequences were available from *C. elegans*, *C. briggsae*, *Drosophila*, and human. Only those genes that did not have a significantly different amino acid substitution rate in the four taxa were used to produce date estimates.

The *C. briggsae*-*C. elegans* sequence data set is the largest available for any pair of congeneric eukaryotes. Such a big sample has a high power for detecting genome-wide trends. For example, the breakpoints of reciprocal translocations and inversions are frequently near repetitive DNA. This has been observed in bacteria (Romero et al., 1999), protozoa (Carlton et al., 2002), yeast (Kellis et al., 2003), insects (Cáceres et al., 1999), mammals (Dehal et al., 2001), and plants (Zhang and Peterson, 1999), but not yet in nematodes. Rearrangements near transposable elements may happen when the element is transposing (Zhang and Peterson, 1999), but most rearrangements are hypothesised to occur by homologous recombination between nontransposing transposable elements, dispersed repeats, or gene family members (Eichler and Sankoff, 2003). We find that translocation and transposition breakpoints are strongly associated with repeats in the *C. elegans* genome.

## 2.2 RESULTS

### 2.2.1 Detection of Conserved Segments and their Length Distribution

Using the BLASTX algorithm (Altschul et al., 1997), we predicted 1784 genes in the 12.9-Mb of *C. briggsae* genomic DNA. The 1784 genes partition the DNA into 756 segments that have been perfectly conserved between the two species. In *C. briggsae*, the segments range from 1 to 19 genes, or 0.6–154 kb. These segments were merged to recreate 252 longer segments that have been fractured by duplications, inversions, or transpositions since speciation. The 252 segments, which we assume to have resulted from fissure of chromosomes by reciprocal translocations, range from 1 to 109 genes in *C. briggsae*, or 1.3–1040 kb (average, 53 kb). In *C. elegans*, the corresponding segments cover 13.7% of the genome, the smallest being one gene (0.4 kb), and the largest 167 genes (954 kb; Figure 2.1 A,B). The segments can be browsed at <http://wolfe.gen.tcd.ie/worm/results.html>. An example of the representation of a conserved segment on the website is shown in Figure 2.2.

If the nine *C. briggsae* supercontigs are concatenated, we have a large 13.3-Mb chunk (the 12.9 Mb sample included internal gaps). If we assume that the 251 translocation breakpoints (and supercontig



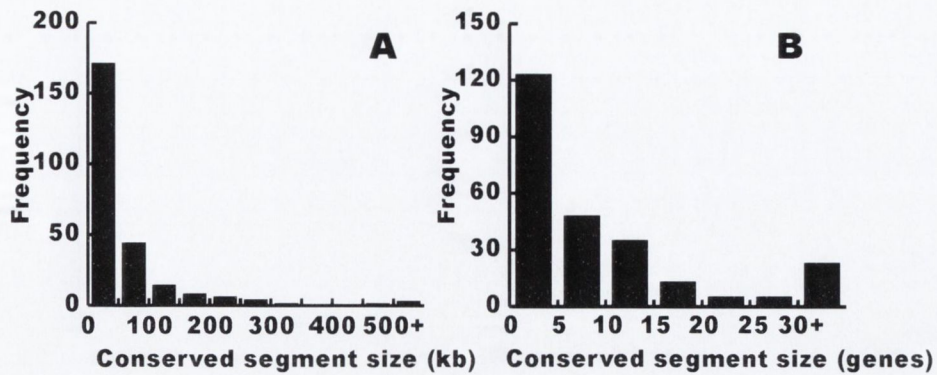


Figure 2.1: Distribution of sizes of conserved segments, measured in units of kilobases (A) and genes (B) with respect to *Caenorhabditis elegans*. These conserved segments were assumed to have resulted from fission of chromosomes by reciprocal translocations.

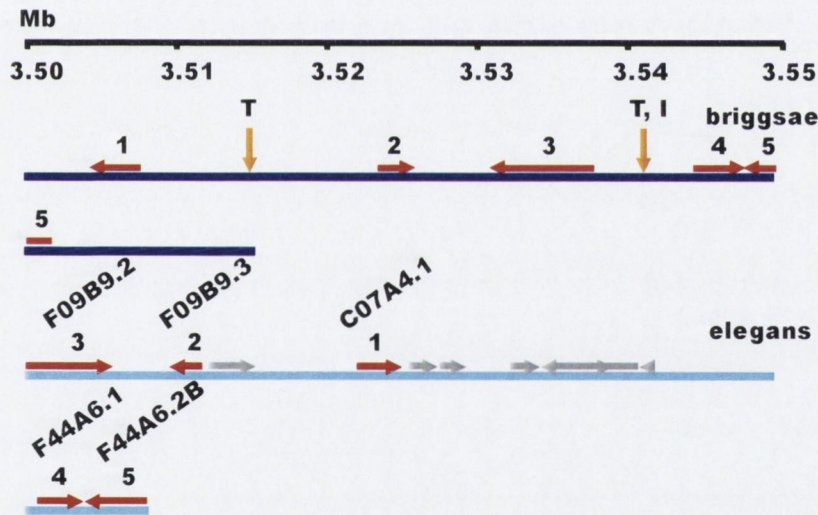


Figure 2.2: The *Caenorhabditis elegans* region surrounding the *sex-1* locus (*F44A6.2B*), and the corresponding region in *Caenorhabditis briggsae*. The pale blue bar wrapped over two lines represents the region between coordinates 10.18–10.23 Mb of *C. elegans* chromosome X, and the navy bar represents 3.50–3.57 Mb of *C. briggsae* contig RWRA. There are five orthologous *briggsae:elegans* genes (red) in the conserved segment, which are identified by the same number (1–5) in the two species, and are named on the *C. elegans* map. Inversion (I) and transposition (T) breakpoints are marked with orange arrows, which are shown arbitrarily on the *C. briggsae* chromosome. A region including three genes (*C07A4.1*, *F09B9.3*, and *F09B9.2*) has been inverted in either *C. elegans* or *C. briggsae* since speciation. Furthermore, a region comprising six genes (gray) between *C07A4.1* and *F44A6.1* in *C. elegans* has transposed to another part of the *C. briggsae* genome, or has transposed into this part of the *C. elegans* genome.



ends) are distributed at random along this chunk, the probability of recovering a segment  $\geq L$  by chance is  $e^{-251L/13.3}$  (Ranz et al., 2001). Of the 252 segments detected, after using the Bonferroni correction for multiple testing, only one is large enough to give a significant result ( $P = 8 \times 10^{-7}$ ). This is a 1.04-Mb segment containing 109 *C. briggsae* genes conserved between *C. briggsae* supercontig FORK and *C. elegans* chromosome X. Gene Ontology classifications are only given in WormBase (<http://www.wormbase.org/>) for 19 of the *C. elegans* orthologues of these 109 *C. briggsae* genes, and there is no obvious relationship between their functions that might provide a selective explanation for why this large segment has been conserved.

### 2.2.2 Differences among and along *C. elegans* Chromosomes

The median length of a conserved segment is significantly greater on the *C. elegans* X chromosome (40.6 kb) than on autosomes (17.0 kb; Mann-Whitney test:  $P < 0.01$ ). It is not known which (if any) of the nine *C. briggsae* supercontigs in the sample originated from its sex chromosome. However, in *C. elegans* sex is determined by counting X chromosomes via X signal elements on the X chromosomes (Akerib and Meyer, 1994). We found the orthologue of the strongest *C. elegans* X signal element, the *sex-1* gene (Carmi et al., 1998), on the *C. briggsae* supercontig RWRA (Figure 2.2). We suggest that RWRA, the largest supercontig (5.0 Mb) in the *C. briggsae* sample, is part of its sex chromosome. RWRA consists of 95 conserved segments matching *C. elegans* autosomes and 23 segments matching the *C. elegans* X chromosome. If RWRA is the *C. briggsae* sex chromosome, the *C. briggsae* sex chromosome must have undergone many reciprocal translocations with autosomes since divergence from *C. elegans*. Conversely, the *C. elegans* X chromosome consists of conserved segments matching five different *C. briggsae* supercontigs, which are unlikely to be all derived from the *C. briggsae* X chromosome.

The 252 conserved segments are scattered over all six *C. elegans* chromosomes (Figure 2.3 A), with 211 being on autosomes and 41 on the X chromosome. Taking Barnes et al.'s (1995) division of *C. elegans* autosomes into arms and centres, we found 102 conserved segments on autosome centres, and 109 on autosome arms (Figure 2.3 A). The median length of a conserved segment was not significantly different among the centres (20.5 kb), the left arms (17.5 kb), and the right arms (15.1 kb) of autosomes (Kruskal-Wallis test:  $P = 0.5$ ).

### 2.2.3 Estimating the *C. briggsae-C. elegans* Divergence Date

Using the divergence of the nematodes from the arthropods at 800–1000 million years ago (Mya; Blaxter, 1998; Brooke, 1999) to calibrate the molecular clock, we estimated the *C. briggsae-C. elegans* divergence date from 92 sets of orthologues. Each set comprised a *C. briggsae* gene, its *C. elegans* orthologue, one or more orthologues from *Drosophila*, and one or more human orthologues. When the nematode-arthropod divergence is taken to be 800 Mya, a 95% confidence interval for the median *C. briggsae-C. elegans* speciation date is 49–94 Mya (median, 70 Mya). If the nematode-arthropod divergence is



taken to be 1000 Mya, the interval becomes 61–118 Mya (median, 88 Mya; Figure 2.4). Our best estimate of the *C. briggsae*-*C. elegans* speciation date is therefore ~50–120 Mya.

## 2.2.4 Duplications

From phylogenetic trees, we identified 27 *C. briggsae* genes that have arisen by 14 duplications from 13 ancestral orthologues at the time of speciation. In 10 of these duplicate pairs, one duplicate has transposed, whereas four of the duplicate pairs have remained adjacent. In two of the four adjacent pairs, one of the duplicates has inverted. Of the 10 duplicates that have transposed, two of the duplicates are on different *C. briggsae* supercontigs. These 10 transpositions and two inversions in *C. briggsae* are the only rearrangements for which we know the genome in which they occurred.

## 2.2.5 Rates of Reciprocal Translocation, Inversion, and Transposition

### *Translocations*

There is no published estimate of the *C. briggsae* genome size, so we assumed that it is about the same size as the *C. elegans* genome (100.1 Mb). To extrapolate from our sample to the entire *C. briggsae* genome, we assumed that the distribution of conserved segment sizes is the same for the unsequenced and sequenced portions. This seems reasonable because the sizes of conserved segments do not differ among autosomes and, although segments from *C. elegans* X are longer than those from autosomes, the fraction of segments from X in our sample (16%) is similar to the fraction of the genome made up by X (18%). Because we found 252 conserved segments in 13% of the *C. briggsae* genome, we estimate that there should be 1953 segments in the entire *C. briggsae* genome. The 1953 conserved segments resulted from the (presumably) six chromosomes present in the last common ancestor (*C. briggsae* has six chromosomes; Nigon and Dougherty, 1949) plus an estimated 1947 breakpoints due to 974 ( $1947/2 = 974$ ) translocations that have occurred since speciation. To calculate the rate of reciprocal translocation, the number of translocations is divided by twice the divergence time (Nadeau and Taylor, 1984). Our estimate of the speciation date, 50–120 Mya, gives a rate of 4.1–9.7 translocations/Myr for the whole genome. Some of our 252 conserved segments consist of only one gene and might have resulted from transpositions; when we include only segments of  $\geq 3$  orthologues, there are 141 conserved segments. Using our 50–120-Mya estimate of the divergence date, this gives a more conservative estimate of 2.3–5.4 translocations/Myr.



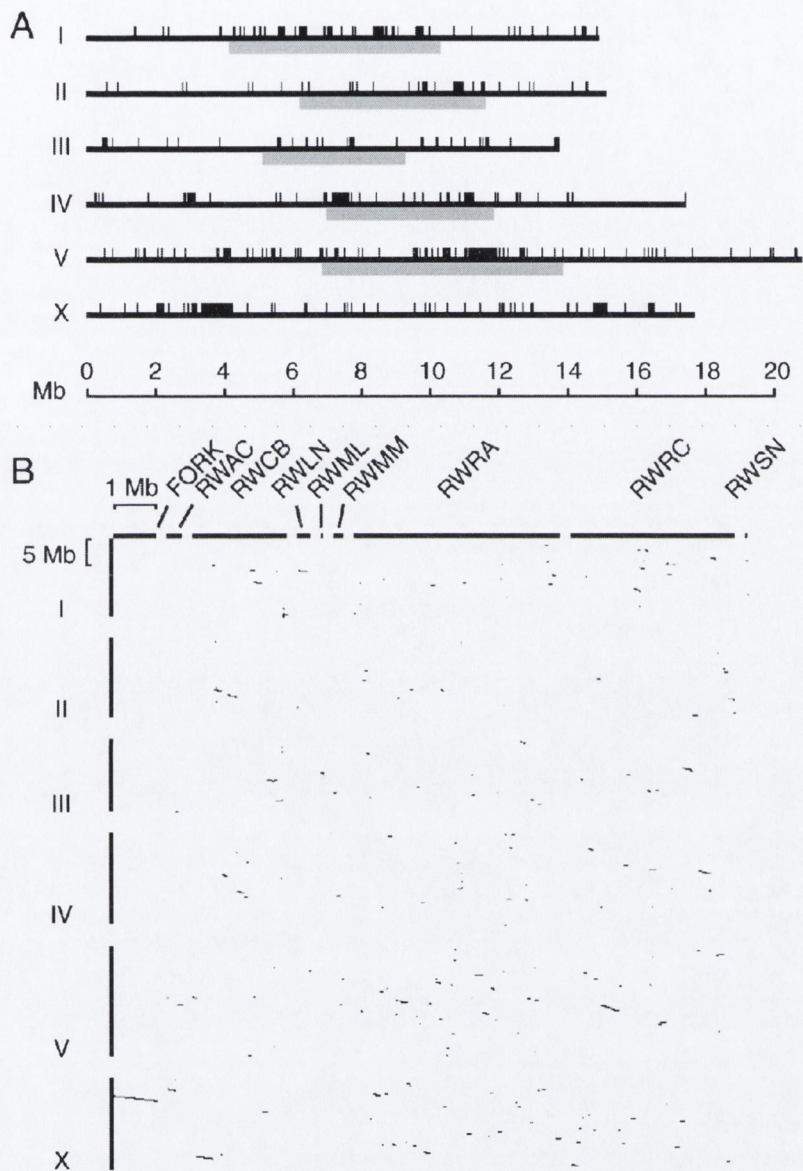


Figure 2.3: (A) Location of conserved segments in the *Caenorhabditis elegans* genome. Gray bars under the autosomes show the “central clusters” described by Barnes et al. (1995). The segments cover ~15% of chromosome I, 7% of II, 10% of III, 13% of IV, 15% of V, and 20% of X. (B) Matrix plot comparison between the *C. elegans* genome (vertical axis) and the nine *Caenorhabditis briggsae* contigs (horizontal axis). Conserved segments are indicated by lines drawn between the positions of the outermost genes in each species.

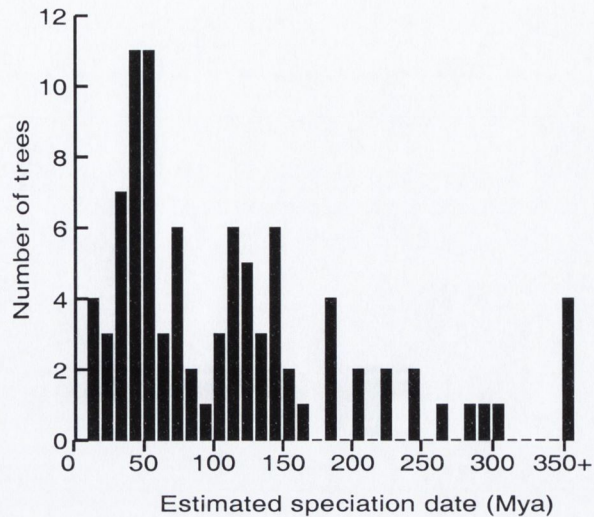


Figure 2.4: Estimates of the *C. briggsae*-*C. elegans* speciation date from 92 sets of *Caenorhabditis briggsae*, *Caenorhabditis elegans*, *Drosophila*, and human orthologues, calculated by taking the nematode-arthropod divergence date to be 1000 Mya.

#### Inversions

We detected 121 inversions, including two inversions of duplicated genes that occurred in *C. briggsae* after speciation, and we estimate that there have been 938 inversions in the two genomes since speciation. Using the same divergence date, this implies a rate of 3.9–9.4 inversions/Myr. In *C. elegans*, the inversions range from 1 to 65 genes, or 0.6–367 kb (median three genes, or 14.4 kb; Figure 2.5 A,B). About two-thirds of the inversions are <25 kb. The autosomes and the sex chromosome do not have a significantly different median inversion breakpoint density in *C. elegans* (Kruskal-Wallis test:  $P = 0.3$ ). Inversion breakpoints are clustered in hotspots on the *C. briggsae* supercontigs: when we concatenate the three largest supercontigs, the median distance between inversion breakpoints is significantly less than would be expected if breakpoints were uniformly distributed (one-sample sign test:  $P = 0.0004$ ). We noticed that next to inversion breakpoints there are often stretches of *C. elegans* genes whose *C. briggsae* orthologues have not been found. We cannot tell whether their *C. briggsae* orthologues have been deleted, or have transposed to or from an as-yet-unsequenced region of the *C. briggsae* genome.

#### Transpositions

We assumed that stretches of *C. elegans* genes whose *C. briggsae* orthologues were not found have resulted from transpositions to or from unsequenced parts of the *C. briggsae* genome (Figure 2.6 A). However, some such transpositions are artifacts. By examining conserved segments, we can see that some of the *C. briggsae* orthologues of *C. elegans* genes have been mistakenly assigned (using BLAST) as the orthologue of a *C. elegans* paralogue. In other cases, the *C. elegans* gene appears to be a misprediction, because it has not any BLAST hit with  $E < 10^{-10}$  in SWISS-PROT or Wormpep. Other such *C. elegans* genes have BLAST hits with  $E < 10^{-25}$  to a neighbouring *C. elegans* gene, and therefore have probably arisen



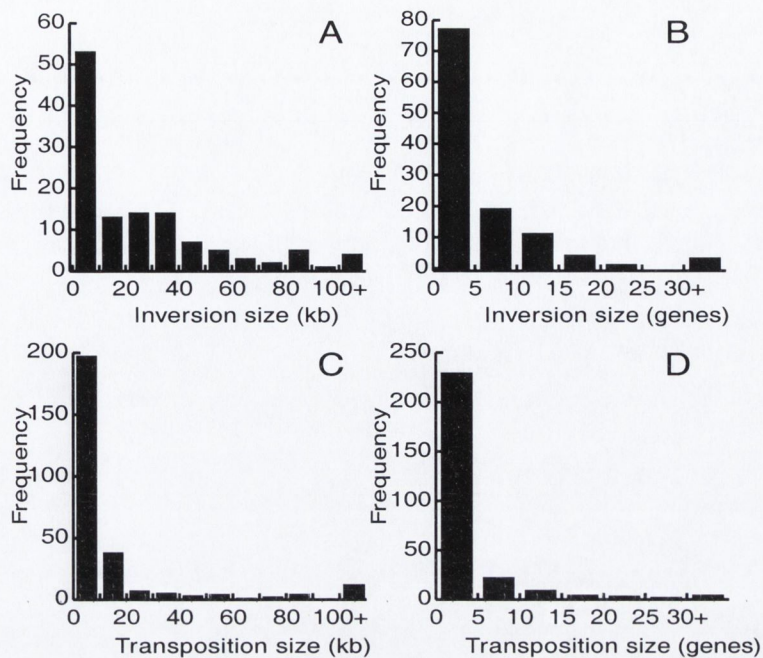


Figure 2.5: (A) Sizes of inversions in kilobases, with respect to *Caenorhabditis elegans*. (B) Sizes of inversions, measures in units of genes. (C) Sizes of transpositions in kilobases. (D) Sizes of transpositions, measured in units of genes.

by tandem duplication since the divergence of *C. elegans* and *C. briggsae*. When we exclude 162 artifacts, 273 transpositions remain. They include 10 transpositions of duplicated genes that have occurred in *C. briggsae*. We estimate that there have been 2116 transpositions in the two genomes since divergence, implying a rate of 8.8–21.2 transpositions/Myr. The 273 transpositions range from 1 to 57 genes, or 0.1–315 kb in *C. elegans* (median, one gene, or 3.3 kb; Figure 2.5 C,D). Most transposed segments of DNA are <30 kb. The size distribution of transpositions differs from that of inversions, being more skewed toward small rearrangements (Figure 2.5 D). For the eight *C. briggsae* duplicate genes that have transposed to the same supercontig, there are 1, 1, 7, 14, 42, 42, 46, and 141 intervening genes, respectively, between their old and new locations. For half of these duplicate pairs, there are  $\leq 20$  genes between the duplicates. Using the method described in Figure 2.6 B, we observed 79 transpositions. If these had all occurred in *C. elegans*, 24 would have been intrachromosomal, and 21 of these 24 to sites >300 genes away. Thus, some intrachromosomal transpositions are probably to sites far away on a chromosome.

#### Overall Rate of Rearrangement

Extrapolating from the sequenced 13% to the entire *C. briggsae* genome, we estimate that 974 reciprocal translocations, 938 inversions, and 2116 transpositions have occurred since speciation. About 4030 chromosomal rearrangements have occurred since divergence of the two species. The ratio of translocations to inversions to transpositions is therefore 1.0:1.0:2.3. Each reciprocal translocation causes two breakpoints, each inversion two breakpoints, and each transposition three breakpoints (Sankoff, 1999). Therefore,



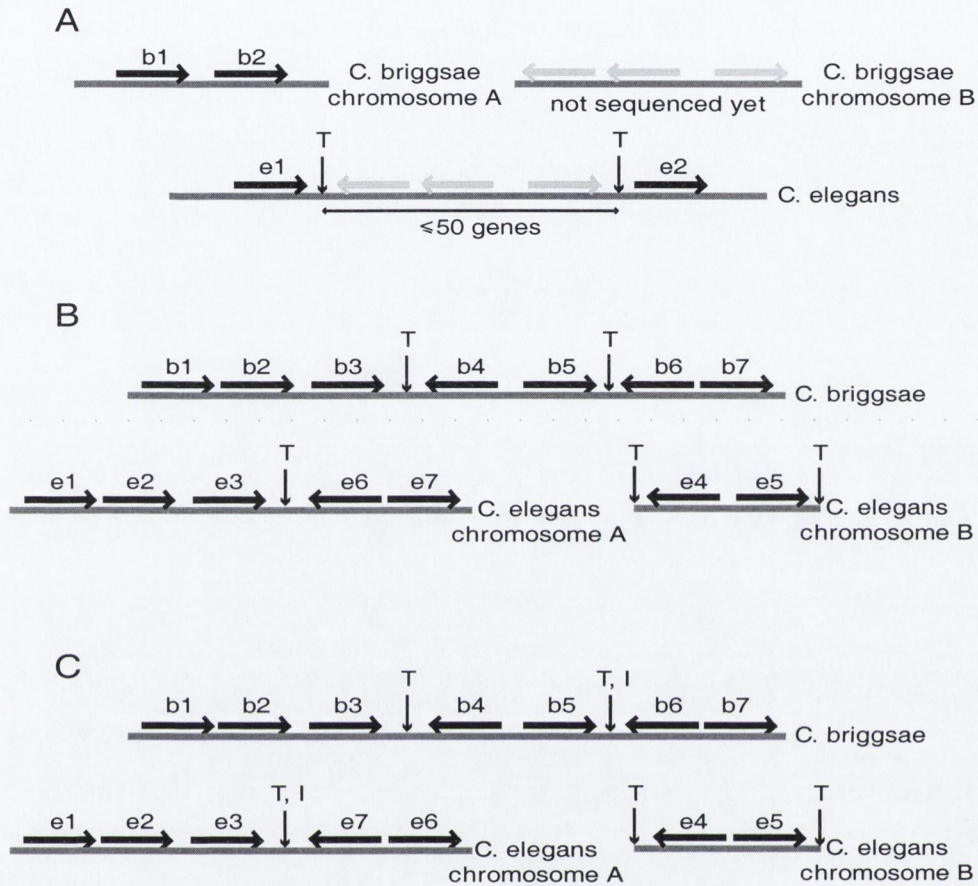


Figure 2.6: Method of detecting inversions and transpositions.

(A) To detect transpositions to or from unsequenced parts of the *Caenorhabditis briggsae* genome, we looked along *C. briggsae* contigs for adjacent genes *b1* and *b2* whose *Caenorhabditis elegans* orthologues *e1* and *e2* are on the same chromosome, where between *e1* and *e2* there are 1–50 *C. elegans* genes with unknown *C. briggsae* orthologues. We assumed that the genes between 1 and 2 have transposed in either *C. briggsae* or *C. elegans*. The symbol T marks transposition breakpoints.

(B) To detect transpositions to or from sequenced parts of the *C. briggsae* genome, we looked along *C. briggsae* contigs for three conserved segments in a row, where in *C. elegans* the first and third segments were close together on the same chromosome, and the middle segment was far away on the same *C. elegans* chromosome or on a different *C. elegans* chromosome. We assumed that the middle segment (genes 4–5) had transposed in either *C. briggsae* or *C. elegans*.

(C) To detect inversions, we looked along *C. briggsae* contigs for three conserved segments in a row, where in *C. elegans* the first and third segments were close together on the same chromosome, and the middle segment was far away on the same *C. elegans* chromosome or on a different *C. elegans* chromosome, and either the first or third segment, or both, had inverted in either *C. briggsae* or *C. elegans*. Here the third segment (genes 6–7) has inverted.



Breakpoint type	Number of Breakpoints	P-value
Translocation	445	$3.3 \times 10^{-5}$
Inversion	185	0.10
Transposition	469	$2.9 \times 10^{-4}$

The number of rearrangement breakpoints in intergenic spacers containing at least one of 33 dispersed repeat families was compared with the number of intergenic spacers in the genome containing one or more dispersed repeats. Only intergenic spacers of 10 kb or shorter were included, of which there are 16,574 in the *Caenorhabditis elegans* genome. The P-values for one-sided  $\chi^2$  tests are given after applying the Bonferroni correction for multiple testing (multiplies the raw P-values by 3).

Table 2.1: Association of Rearrangement Breakpoints with Repeats.

there have been  $\sim 10,200$  chromosome breakages since speciation, which is 5100 breakages per species, or  $\sim 51$  breakages/Mb. Using our 50–120-Mya divergence date, this implies a rate of 42–102 breakages/Myr, or 0.4–1.0 breakages/Mb per Myr.

## 2.2.6 Association of Breakpoints with Repetitive DNA

We obtained the distribution of 33 dispersed repeat sequences in the *C. elegans* genome from WormBase ([http://www.sanger.ac.uk/Projects/C.elegans/WORMBASE/GFF\\_files.shtml](http://www.sanger.ac.uk/Projects/C.elegans/WORMBASE/GFF_files.shtml); Stein et al., 2001).

When we pool all 33 repeats, there is a significant association between dispersed repeats and both translocation and transposition breakpoints in *C. elegans* (Table 2.1). However, no association is seen for inversion breakpoints. For two individual dispersed repeats, the association with transposition breakpoints is significant ( $P < 0.05$ ; Table 2.2): CeRep20 and CeRep37. However, the significance of the association is marginal for CeRep20 ( $P = 0.045$ ), whereas the small sample size for CeRep37 makes the test result unreliable.

Translocation breakpoints tend to be next to four different repeats: CeRep13, CeRep15, CeRep19, and CeRep32. *C. elegans* has compound repeats, listed on the Sanger Institute web site (<http://www.sanger.ac.uk/Projects/C.elegans/repeats/>). The only one associated with translocation breakpoints is CeRep13-CeRep18-CeRep18-CeRep33-CeRep18-CeRep13 ( $P = 0.01$ ; Table 2.2). However, this is simply owing to the association of CeRep13 with breakpoints, because breakpoints are often near CeRep13/CeRep18/CeRep33, but not CeRep13 + CeRep18 + CeRep33. The association of CeRep19 and CeRep32 with translocation breakpoints is marginally significant ( $P \leq 0.05$ ), but that of CeRep13 and CeRep15 is strong ( $P \leq 0.005$ ). CeRep13 is a 26-bp sequence that is repeated  $\sim 1350$  times in the *C. elegans* genome, whereas CeRep15 is a 63-bp sequence of which there are about 910 copies. Both these repeats seem to be derived from transposable elements. CeRep13 is 96% identical over 24 bp to the 24-bp terminal inverted repeat (TIR) of *Cele11*, which is thought to be a nonautonomous relative of Tc2, a Tc1/*mariner* family



Dispersed repeat	All 16,574 spacers	Translocation breakpoints	P-value for translocations	Transposition breakpoints	P-value for transpositions
CeRep10	582	23	1.00	25	1.00
CeRep11	137	4	1.00	7	1.00
CeRep12	553	17	1.00	24	1.00
CeRep13	354	22	<b>0.003</b>	15	1.00
CeRep14	320	16	0.44	11	1.00
CeRep15	186	14	<b>0.005</b>	10	1.00
CeRep17	345	19	0.06	11	1.00
CeRep18	197	10	1.00	11	0.90
CeRep19	685	33	<b>0.02</b>	18	1.00
CeRep20	144	9	0.47	11	<b>0.045</b>
CeRep21	177	8	1.00	11	0.36
CeRep22	122	6	1.00	8	0.75
CeRep23	708	27	1.00	31	0.42
CeRep24	625	20	1.00	23	1.00
CeRep25	7	1	1.00	0	1.00
CeRep26	154	4	1.00	7	1.00
CeRep27	71	5	1.00	4	1.00
CeRep28	92	4	1.00	5	1.00
CeRep29	150	6	1.00	5	1.00
CeRep30	37	2	1.00	4	0.50
CeRep31	23	1	1.00	2	1.00
CeRep32	226	15	<b>0.02</b>	6	1.00
CeRep33	22	1	1.00	1	1.00
CeRep34	321	10	1.00	16	0.78
CeRep35	177	9	1.00	5	1.00
CeRep36	187	6	1.00	6	1.00
CeRep37	122	4	1.00	11	<b>0.006</b>
CeRep38	310	12	1.00	15	1.00
CeRep39	14	1	1.00	1	1.00
CeRep40	122	5	1.00	2	1.00
CeRep41	49	3	1.00	1	1.00
CeRep42	110	5	1.00	5	1.00
CeRep43	590	27	0.17	23	1.00
29+35+36+40	24	2	1.00	0	1.00
29/35/36/40	391	14	1.00	10	1.00
17+19+32	166	11	0.12	5	1.00
17/19/32	720	33	0.06	18	1.00
13+18+33	17	1	1.00	1	1.00
13/18/33	383	22	<b>0.01</b>	15	1.00
34+43	212	10	1.00	7	1.00
34/43	699	27	1.00	29	1.00
24+38	308	12	1.00	15	1.00
24/38	627	20	1.00	23	1.00

The number of translocation/transposition breakpoints in intergenic spacers containing a particular dispersed repeat was compared with the number of intergenic spacers in the genome containing that repeat. Only intergenic spacers of 10 kb or shorter were included, of which there are 16,574 in the *Caenorhabditis elegans* genome. We tested whether breakpoints are associated with five compound repeats. For example, for the compound repeat CeRep19-CeRep32-CeRep17-CeRep19, we tested whether intergenic spacers containing breakpoints tend to contain all members of the repeat (17 + 19 + 32), or at least one member of this repeat (17/19/32). The P-values for one-sided  $\chi^2$  tests are given after applying the Bonferroni correction for multiple testing (multiplies the raw P-values by 43).

Table 2.2: Association of Translocation and Transposition Breakpoints with Particular Repeats.



transposon (Oosumi et al., 1996). CeRep15 is 89% identical over 63 bp to part of the 170-bp TIR of *Cele7*, also thought to be a nonautonomous DNA transposon (Oosumi et al., 1995). We searched the *C. briggsae* genomic DNA for CeRep13 and CeRep15 using FASTA (Pearson and Lipman, 1988). Homologues of CeRep13 seem to be present in the *C. briggsae* genome, because it has hits of 91% identity over 22 bp.

It is possible that rearrangement breakpoints could be associated with repeated gene sequences. To investigate this, we used BLASTP (Altschul et al., 1997) with an *E*-value cutoff of  $10^{-100}$  to define families of highly similar genes. There are 1252 families, containing 3901 genes. The proportion of translocation breakpoints that have a gene family member on one or both sides (41%) is significantly greater than the proportion of all *C. elegans* intergenic spacers having a family member on one or both sides (33%; one-sided  $\chi^2$  test;  $P = 0.0001$ ). A strong association is also seen for transposition breakpoints (one-sided  $\chi^2$  test;  $P = 0.0002$ ), but none for inversion breakpoints.

## 2.3 DISCUSSION

The average size of a conserved segment is 53 kb in *C. briggsae*. This is much larger than the 8.6-kb average found by Kent and Zahler (2000), even though they analysed a subset of the same *C. briggsae* sequences (8.1 Mb of 12.9 Mb). There are three reasons for the difference. First, Kent and Zahler did not realise that the order and spacing of clones along *C. briggsae* chromosomes are known. The average size of the clones in their sample was 36 kb, whereas the average size of the supercontigs in our sample is 1486 kb. They underestimated the average size of a conserved segment because many clones end before the segment ends. Second, because their method allowed up to 50 kb of contiguous nonsyntenous DNA within a conserved segment in *C. elegans* but only up to 1 kb in *C. briggsae*, it was biased toward finding shorter conserved regions in *C. briggsae* than *C. elegans*. Third, instead of their approach of defining conserved segments by an arbitrary gap size, we strove for a more biologically meaningful approach by searching for the fragments into which chromosomes have been splintered by translocations. We followed Sankoff's (1999) suggestion and regarded inversions and transpositions within translocated segments as noise. For example, Kent and Zahler split the chromosomal region containing the *sex-1* locus into nine segments, partitioning the DNA at poorly conserved noncoding stretches or where there have been small inversions and transpositions. In contrast, we found one large conserved segment in the *sex-1* region (Figure 2.2).

Since publication of this work (Coghlan and Wolfe, 2002), the entire *C. briggsae* genome has been sequenced. Dr. Lincoln Stein and Dr. Todd Harris at Cold Spring Harbor have identified ~4800 conserved segments between the entire *C. briggsae* and *C. elegans* genomes, with a mean size of 37.5 kb (Stein et al. 2003; note that  $4800 \times 37.5 \text{ kb} \approx 180 \text{ Mb}$ , which is larger than the ~100 Mb genome size because some of the segments overlap). They found more and smaller segments than we expected based on 13% of the *C. briggsae* genome (~1900 segments). This is probably because our conserved segments had to contain



at least one gene, while those of Stein et al. (2003) were based on nucleotide-level alignments, so included shorter regions of conserved noncoding DNA. Furthermore, they included overlapping segments (we did not), which will have increased their count of segments.

Despite the huge amount of genome rearrangement since *C. elegans* diverged from *C. briggsae*, both species have six chromosomes (Nigon and Dougherty, 1949). In the clade of nematodes to which *Caenorhabditis* belongs (clade V; Figure 1.3), which arose ~550 Mya (Vanfleteren et al., 1994), most species have a haploid chromosome number of  $n = 5-6$  (Blaxter, 2000). This contrasts with the frequent chromosome fissions and fusions that have occurred in ~100 Myr of primate evolution (Haig, 1999), suggesting that there may be selection for a stable number of chromosomes in clade V nematodes.

Kent and Zahler (2000) found that 63 of their 100 longest conserved segments were near the middle of *C. elegans* autosomes and surmised that “chromosome arms appear to be more susceptible to rearrangement.” We found no significant difference between the lengths of conserved segments in *C. elegans* autosome arms and centres. However, in their recent comparison of the entire *C. briggsae* and *C. elegans* genomes, Stein et al. (2003) found that conserved segments are longer in the centres of *C. elegans* chromosomes (mean 44 kb) than in the arms (mean 26 kb). Stein et al. (2003) hypothesise that there are more rearrangements in the arms because the arms are more repeat-rich than the centres, and repeats promote ectopic recombination.

We found a large difference between the median segment size on X (41 kb) and on autosomes (17 kb). This was confirmed for the whole *C. briggsae* genome by Stein et al. (2003), who found that the X chromosome has undergone less rearrangements (31 rearrangements/Mb per Myr) than the autosomes (52 rearrangements/Mb per Myr). Both interchromosomal and intrachromosomal rearrangements are less frequent on the X than autosomes (see Table 9 in Stein et al., 2003). The difference seems far too large to be attributable to a lower sensitivity for detecting conserved segments in gene-poor regions like the X chromosome. Rather, X appears to be better conserved than the autosomes, which must be caused by a lower rate of occurrence or fixation of rearrangements of X. There may be fewer X rearrangements than autosomal rearrangements because of the lower density of some repeats on the *C. elegans* X chromosome (The *C. elegans* Sequencing Consortium, 1998; Surzycki and Belknap, 2000). Alternatively, the rate of fixation of rearrangements may be different for X chromosomes and autosomes. Ohno (1967) hypothesised that in species such as *C. elegans* that have dosage compensation systems in which X genes in XX organisms are down-regulated, X-autosomal translocations will be more deleterious than autosome-autosome translocations. Furthermore, if most rearrangements are deleterious recessive, for example, because they upset regulation of expression, we would expect X rearrangements to be fixed less often than autosomal rearrangements, because selection against deleterious recessive mutations is stronger on the X than autosomes (Charlesworth et al., 1987). On the other hand, if most rearrangements are selectively neutral, X may have a lower fixation rate because of a lower susceptibility to hitchhiking



effects compared with the recombinationally quieter centres of autosomes (Barnes et al., 1995). A further possibility is that there is selection against rearrangements of the region(s) of the *C. elegans* X chromosome from which dosage compensation is initiated, as is seen in mammals (Nesterova et al., 1998).

Translocation and transposition breakpoints are often near repetitive DNA in the *C. elegans* genome, such as gene family members and dispersed repeats. Ectopic recombination between repeats may cause reciprocal translocations. Further study is needed to find out why transposition breakpoints tend to be near dispersed repeats (Table 2.1) and gene family members. It is possible that we have sometimes mistaken two translocations that were between sites close to each other on the same pair of chromosomes as a transposition.

When counting rearrangements, we could detect only inversions or transpositions of genes within conserved segments. As a result, some transpositions may have been mistaken for translocations, for example, a three-gene segment that transposed to a position between conserved segments. Furthermore, we may not have detected all inversions and transpositions, for example, if an entire conserved segment was inverted. Many rearrangement breakpoints have been reused in this way since mouse and human diverged (Pevzner and Tesler, 2003b). Another possible source of error is that we assumed that stretches of *C. elegans* genes whose *C. briggsae* orthologues have not been found were caused by transpositions to or from an as-yet-unsequenced region of the *C. briggsae* genome (Figure 2.6 A), but it could be that the *C. briggsae* orthologues have been deleted. Our count of rearrangements may also have been affected by problems that are not specific to our method. First, the average size of a *C. briggsae* supercontig in our sample was 1486 kb, so we may not have detected rearrangements  $> 1.5$  Mb. Second, rearrangements that occur twice cannot be detected (Sankoff, 1999). Third, it can be impossible to distinguish between three overlapping inversions and a single transposition (Blanchette et al., 1996). Following Nadeau and Taylor (1984), we attributed the few such ambiguous cases to inversions. However, some such inversions may have been in fact transpositions, because we found that transpositions are more common than inversions in *Caenorhabditis*. Fourth, we could not tell a reciprocal translocation apart from a chromosome fusion followed by a fission unless both of the translocation breakpoints had been found. We assumed ambiguous cases to be reciprocal translocations, not chromosome fusions or fissions, because both species have six chromosomes (Nigon and Dougherty, 1949). Thus, we will not have detected if a fusion was followed by a fission in one of the species, or if a chromosome fission occurred in both species since divergence. Our estimate of  $\sim 4000$  rearrangements does not seem to have been affected much by the difficulties due to having a partial genome sequence, because it agrees well with the later estimate of  $\sim 4400$  rearrangements based on the whole *C. briggsae* genome (Stein et al., 2003).

We estimated that *Caenorhabditis* has a rearrangement rate of 0.4–1.0 breakages/Mb per Myr. When Stein et al. (2003) later repeated this analysis based on the whole *C. briggsae* genome, they estimated a similar value of 0.5–0.7 breakages/Mb per Myr. This is  $\sim 30$ –50 times the mammalian rate observed



by Pevzner and Tesler (2003a). Moreover, the nematode rate is  $\sim 5$ – $35$  times faster than the rate in *Drosophila*, previously reported to be the fastest rate among eukaryotes (0.02–0.09 breakpoints/Mb per Myr; González et al., 2002). The high rate in *Drosophila* is paralleled by that in *Anopheles* (0.04–0.07 breakpoints/Mb per Myr; Sharakhov et al., 2002). Error in the estimated *C. briggsae*-*C. elegans* divergence date would make our rate estimate inaccurate, but it seems unlikely that we have overestimated the rate of rearrangement. For nematodes to have the same rearrangement rate as *Drosophila*, the *C. briggsae*-*C. elegans* divergence date would have to be  $> 570$  Mya; however, the nematode order to which *Caenorhabditis* belongs arose only  $\sim 400$  Mya (Vanfleteren et al., 1994). *Caenorhabditis* and *Drosophila* differ not only in the rate, but also in the type, of rearrangement seen. In *Caenorhabditis*, translocations and inversions are roughly equally frequent, inversions being slightly more common (Stein et al., 2003). Likewise, in mammals, small inversions are far more frequent than translocations (Pevzner and Tesler, 2003a). In contrast, in arthropods translocations are very rare compared to inversions (González et al., 2002; Sharakhov et al., 2002). The rate of gene transposition is also an order of magnitude less frequent in *Drosophila* than in *Caenorhabditis* (Ranz et al., 2003).

González et al. (2002) analysed in situ hybridisation data from three *Drosophila melanogaster* chromosomes and the corresponding *Drosophila repleta* chromosomes, and used a maximum likelihood method to estimate the number of inversions that have occurred since the divergence of the homologous *D. melanogaster*-*D. repleta* chromosomes. Their likelihood method was designed to give an unbiased estimate of the number of rearrangements; thus differences between our *Caenorhabditis* results and their *Drosophila* results are probably not caused by differences between the methods used. However, some differences between the results are probably due to differences in data quality. For example, it is likely that they have underestimated the rate of small rearrangements in *Drosophila* for two reasons. First, because the orientation of the *Drosophila* markers was not known in both species, they could not detect inversions of single markers (for comparison,  $\sim 40\%$  of the *Caenorhabditis* inversions we detected were one gene long; Figure 2.5 B). Second, their physical map only had one marker per 49 kb in its densest regions, thus the smallest inversion that they could detect was  $\sim 100$  kb long (for comparison,  $\sim 95\%$  of the *Caenorhabditis* inversions detected were  $< 100$  kb long; Figure 2.5 A). Zdobnov et al. (2002) identified many small inversions by comparing the whole genome sequences of *D. melanogaster* and *Anopheles gambiae*, but unfortunately they did not estimate the rate of small rearrangements in arthropods. In contrast to the case for small rearrangements, González et al. (2002) will have detected more long inversions than we did, because the average size of a *C. briggsae* supercontig in our sample was  $\sim 1.5$  Mb, whereas their markers spanned whole chromosomes (each  $> 20$  Mb).

We suggest four reasons why *Caenorhabditis* chromosomes may have a faster rearrangement rate than those of *Drosophila*. First, the generation time of *Caenorhabditis* is 4–5 times shorter (3–4 days for *C. elegans*, compared with  $\sim 2$  weeks for *Drosophila*). Second, *C. elegans* and *C. briggsae* may have a



smaller effective population size than *Drosophila*, because they are largely self-fertilising but *Drosophila* is not (Sivasundar and Hey, 2003). Third, *C. elegans* chromosomes may be more prone to hitchhiking effects than those of *Drosophila*, because in *C. elegans* the more gene-rich regions of autosomes have the lowest recombination rates, but the opposite is true for *Drosophila* (Barnes et al., 1995). In other words, if a selectively neutral rearrangement occurs near a positively selected gene, in *C. elegans* it is less likely to be separated from the selected gene by meiotic recombination, and so is more likely to undergo a selective sweep with that gene. These three reasons may also lie behind the faster substitution rate in *C. elegans* compared with *Drosophila*. However, the amino acid substitution rate is usually less than two times faster in a *C. elegans* gene than in its *Drosophila* orthologue (see Figure 3 in Mushegian et al., 1998), whereas the rearrangement rate is at least four times faster in *C. elegans*. Our fourth reason is the only one that may contribute to a higher rearrangement rate in *C. elegans* but not to a higher substitution rate. It is that in selfing species like *C. elegans* and *C. briggsae*, rearrangements that are deleterious when heterozygous are more likely to persist than in an out-crossing species, because homozygous individuals arise sooner (Lande, 1979). If this is true, we would expect non-selfing species of *Caenorhabditis*, such as *Caenorhabditis remanei* (Baird et al., 1992), to have a lower rate of rearrangement compared with *Drosophila* than do *C. elegans* and *C. briggsae*. We would also expect greater karyotype variability in *C. elegans* populations than in *Drosophila* or *C. remanei* populations. Genomic sequence from non-selfing *Caenorhabditis* species and data on the karyotype variability in wild *C. elegans* populations could provide clues as to why there is a rate difference.

Do all nematodes have high rates of chromosomal rearrangement? In a region sequenced from *Pristionchus pacificus* (a clade V diplogasterid; Figure 1.3), only 3/10 adjacent gene-pairs (30%) are conserved in *C. elegans*, scattered along one *C. elegans* chromosome (Lee et al., 2003). Surprisingly, there is about the same degree of synteny conservation between *C. elegans* and its more distant relative the filarial nematode *Brugia malayi*: in an 11-gene region compared, 4/10 adjacent gene-pairs (40%) are conserved (Guiliano et al., 2002). We would expect a faster rate of evolution in largely self-fertilising species with short 3–4 day generation times such as *C. elegans* and *P. pacificus*, than in an out-crossing species like *B. malayi* with a life cycle lasting > 3 months. Thus, more rearrangements may have occurred in the *Pristionchus* and *Caenorhabditis* lineages than in the *Brugia* lineage.

## 2.4 FUTURE WORK

The now fully sequenced *C. briggsae* genome sequence (Stein et al., 2003) and the near-finished *Brugia malayi* genome sequence (see Chapter 1) will provide a perfect opportunity to further investigate evolution of chromosome structure in nematodes, for example:

- Gene transposition is very frequent in nematodes compared to arthropods (Ranz et al., 2003). It is of interest to test whether genes of repetitive nature tend to be transposed in nematodes, as in



*Drosophila* (Ranz et al., 2003). Furthermore, it is important to determine the most common molecular mechanism by which genes transpose in nematodes, of the three proposed mechanisms (Ranz et al., 2003).

- We would like to identify *C. briggsae*-*C. elegans* segments and *Caenorhabditis*-*Brugia* segments that are likely to have been conserved by natural selection rather than by chance. A vital task will be to distinguish segments that have been conserved because they contain operons (Stein et al., 2003), from those containing non-operonic co-regulated genes (Roy et al., 2002; Lercher et al., 2003; Lee and Sonnhammer, 2003).
- Little is known about the birth and death of operons. Of *C. elegans* operons, 96% are conserved in *C. briggsae* (Stein et al., 2003) and at least one is conserved in the closely related rhabditid *Dolichocephalus* (Evans et al., 1997). However, operon structure across the Phylum Nematoda seems to be in flux — the one operon analysed so far in *Pristionchus pacificus* (a clade V diplogasterid; Figure 1.3) is not conserved in *C. elegans* (Lee and Sommer, 2003). One informative path is to identify *C. elegans* operons that have been conserved or broken in *Brugia*, which seems to also have operons (Takacs et al., 1988).
- One of the great mysteries of *C. elegans* chromosome evolution is the difference in evolutionary rate between the arms and centres of autosomes. Nobody knows whether this pattern is present in other nematodes. The *B. malayi* Sequencing Project aims to reach 8X coverage by October 2004, a level of coverage that should provide clearly assembled chromosomes (E. Ghedin, pers. comm.). It will then be exciting to test whether arm-centre differences are seen in *B. malayi* chromosomes.

## 2.5 METHODS

### 2.5.1 Sources of Sequence Data

Nine supercontig DNA sequences from the *C. briggsae* Sequencing Project at the Washington University Genome Sequencing Center (<http://genome.wustl.edu/gsc/>) generated from a fingerprint map of *C. briggsae* (M. Marra, J. Schein, and R. Waterston, unpubl.) were downloaded from the WormBase site (<http://www.wormbase.org/>; Stein et al., 2001) in July 2001. The *C. briggsae* data consist mainly of genes requested by the Worm Community to be sequenced (Baillie and Rose, 2000) and are therefore not a random sample of the genome. The nine supercontigs range from 70 to 5015 kb. Because some of these supercontigs contained large internal gaps, we subdivided supercontigs at any internal gap of > 2 kb. The resulting 20 contigs range from 51 to 2288 kb (median, 369 kb) and totalled 12.9 Mb. The 19,957 *C. elegans* protein sequences from Wormpep54 were downloaded from [http://www.sanger.ac.uk/Projects/C\\_elegans/wormpep/](http://www.sanger.ac.uk/Projects/C_elegans/wormpep/) in July 2001. We discarded 586 Wormpep proteins from the genes of transposable elements and genes similar to transposable element genes, 31 from



genes whose chromosomal coordinates in *C. elegans* are unknown, and 713 from alternatively spliced genes (retaining the longest splice variant only); 18,627 proteins remained. *C. elegans* gene coordinates corresponding to ACeDB release WS44 were downloaded in July 2001 from [http://www.sanger.ac.uk/Projects/C\\_elegans/WORMBASE/GFF\\_files.shtml](http://www.sanger.ac.uk/Projects/C_elegans/WORMBASE/GFF_files.shtml).

### 2.5.2 Predicting *C. briggsae* Genes

The *C. briggsae* contigs were largely unannotated, so we predicted *C. briggsae* genes using a spliced alignment approach similar to that of Mironov et al. (1998). This was feasible because protein coding regions are well conserved between the two species, but intergenic regions and introns are not (Kent and Zahler, 2000). Regions of the *C. briggsae* contigs homologous to *C. elegans* proteins were identified using BLASTX (Altschul et al., 1997) with the BLOSUM62 scoring matrix (Henikoff and Henikoff, 1992), using the SEG filter (Wootton and Federhen, 1996), and storing all hits with an *E*-value of  $\leq 0.1$ . There were 99,221 BLASTX hits. Because BLASTX does not always accurately distinguish between orthologues and paralogues, we kept any overlapping hits having *E*-values within a factor of 90 of each other. Nearby BLASTX hits to the same *C. elegans* protein were assumed to correspond to the exons of a *C. briggsae* homologue, and were merged so long as they were on the same strand of the *C. briggsae* contig. To avoid merging hits that were implausibly far apart on a *C. briggsae* contig, any *C. briggsae* intron could not be  $> 7700$  bp, and the summed length of introns in a *C. briggsae* gene could not exceed 8150 bp. These numbers (90, 7700, and 8150) were chosen on inspection of the results. To prevent mistaken merging of tandemly repeated genes on a *C. briggsae* contig, the following rule was used, where "left" and "right" refer to the position on the *C. briggsae* contig. The left BLASTX hit had to start in the *C. elegans* protein before the right hit ended in the *C. elegans* protein, and the left hit had to end in the *C. elegans* protein before the right hit started in the *C. elegans* protein, or the hits overlap by  $< 1100$  amino acids. After merging BLASTX hits to predict genes, we found a nonoverlapping set of the most significant *C. briggsae* genes along each supercontig. Lastly, *C. briggsae* genes that hit  $< 45\%$  of the length of the *C. elegans* protein, or had BLASTX *E*-values of  $\geq 10^{-5}$ , were deleted, as they were probably pseudogenes. On the nine supercontigs, we predicted 1934 *C. briggsae* genes. We will not have detected *C. briggsae* genes that do not have homologues in *C. elegans*.

### 2.5.3 Finding Orthologues

We did not use synteny data to define orthologues, only sequence identity and phylogenetic trees, because we wanted to use orthologues to gauge synteny conservation. The 1934 *C. briggsae* genes hit 1804 different *C. elegans* proteins in BLASTX. If a *C. briggsae* gene hit only one *C. elegans* protein, then the *C. briggsae* and *C. elegans* genes were taken to be one-to-one orthologues. Based on BLASTX results alone, 1704 one-to-one orthologue pairs were found. Some of these orthologous pairs were detected from BLASTX hits having *E*-values as high as  $10^{-6}$ . For the remaining 230 *C. briggsae* genes, it was necessary to draw



151 different phylogenetic trees to deduce orthology. To find an outgroup for a tree of a *C. briggsae* gene and its *C. elegans* hits we used BLASTP (Altschul et al., 1997) with an *E*-value cutoff of  $\leq 0.1$  to compare the *C. elegans* hits to Wormpep54 (19,957 proteins) and to SWISS-PROT (July 2001). For each *C. elegans* protein in the tree, the outgroup was either the top-scoring *C. elegans* hit for which a *C. briggsae* orthologue had previously been identified from BLASTX results, or the top-scoring non-*C. elegans* hit, whichever had the highest score. The sequences for a tree were aligned using CLUSTALW (Thompson et al., 1994), and a maximum parsimony phylogenetic tree was drawn using protpars (Felsenstein, 1993). We bootstrapped the trees using 1000 bootstrap replications in the seqboot algorithm (Felsenstein, 1993). Only nodes with bootstraps of  $\geq 80\%$  were used to deduce orthology.

Our final *C. briggsae* data set contains 1934 genes: 1774 genes in one-*C. briggsae*-to-one-*C. elegans* orthology relationships, and 190 other genes in the following relationships:

- (1) 13 genes in one-*C. briggsae*-to-many-*C. elegans* relationships;
- (2) 46 genes in many-*C. briggsae*-to-one-*C. elegans* relationships: 23 for which the *C. elegans* orthologue is known and 23 for which the *C. elegans* orthologue is unresolved;
- (3) 4 genes in many-*C. briggsae*-to-many-*C. elegans* orthology relationships;
- (4) 13 genes whose *C. elegans* orthologue has been deleted since speciation or had not been sequenced yet; and
- (5) 114 remaining *C. briggsae* genes whose orthology is unresolved.

For 137 of the genes, orthology could not be decided owing to lack of a suitable outgroup or low bootstraps in trees.

The 137 *C. briggsae* genes of unresolved orthology (mainly histone genes) and the 13 with deleted orthologues were ignored in the subsequent analysis, leaving 1784 *C. briggsae* genes that hit 1792 different *C. elegans* genes, of which 1744 were one-to-one orthologues.

#### 2.5.4 Estimating the *C. briggsae*-*C. elegans* Divergence Date

We downloaded 161,296 human proteins and 35,108 *Drosophila* proteins from GenBank (<http://www.ncbi.nlm.nih.gov/Entrez>; December 2001). To find *C. elegans* orthologues of these proteins, we compared them with Wormpep using BLASTP (Altschul et al., 1997) with the SEG filter (Wootton and Federhen, 1996). If a human protein hit a *C. elegans* protein with a BLASTP *E*-value of  $< 10^{-20}$ , and the *C. elegans* protein with the second strongest hit had an *E*-value that differed by a factor of  $10^{20}$  or more, then the *C. elegans* protein was considered to be the orthologue of the human protein. We found 238 sets of orthologues, each set containing a *C. briggsae* gene, its *C. elegans* orthologue, one or more human orthologues, and one or more *Drosophila* orthologues. For each set, we aligned the proteins using CLUSTALW (Thompson et al., 1994), and made a guide-tree using protdist and neighbor from the PHYLIP package (Felsenstein, 1993). We discarded 33 orthologue sets for which the human sequences



did not group together and/or the *Drosophila* sequences did not group together, leaving 205 sets. For each orthologue set, the alignment and guide-tree were used as input for Gu and Zhang's (1997) program GAMMA, which estimated an  $\alpha$  parameter for the  $\Gamma$  distribution used to correct for rate variation among amino acid sites. For 31 trees, GAMMA could not estimate the  $\alpha$  parameter. For the remaining 174 trees, we used the two-cluster test (Takezaki et al., 1995) to check for unequal rates between lineages, taking human to be the outgroup to *Drosophila* and *Caenorhabditis* (Aguinaldo et al., 1997); 92 trees passed the test at the 5% significance level. For each tree, the branch lengths were re-estimated under the assumption of rate constancy, using Takezaki and Nei's (1995) program with the  $\Gamma$  correction for multiple hits. Although the exact branching order of the chordates, arthropods, and nematodes continues to be hotly debated (Mushegian et al., 1998; Wang et al., 1999), most estimates of the divergence of these three phyla range from 800 to 1000 Mya (Blaxter, 1998; Brooke, 1999). We calibrated the linearised trees by taking the nematode-arthropod divergence to be 800–1000 Mya.

### 2.5.5 Finding Conserved Segments and Classifying Breakpoints by Mutation Type

When two species are compared, any region of their genomes in which gene content and order are conserved is a "conserved segment" (Sankoff, 1999). Between two adjacent conserved segments is a "breakpoint" (Sankoff, 1999) caused by translocation, inversion, duplication, or transposition. We searched for all perfectly conserved segments on the *C. briggsae* supercontigs: segments in which gene order and orientation are perfectly conserved with *C. elegans*. To estimate the size distribution of different types of mutation, the breakpoints within *C. briggsae* contigs were classified as duplication, translocation, inversion, or transposition breakpoints as described below.

From phylogenetic trees, we identified *C. briggsae* genes that have arisen by duplication since speciation. If two *C. briggsae* duplicates that arose from one orthologue were adjacent, we called the breakpoint between them a duplication breakpoint; if one of the duplicates is inverted, it is also an inversion breakpoint. These breakpoints were subsequently ignored, thereby enlarging the original conserved segments. A conserved segment was then taken to be the region between two as-yet-unexplained breakpoints. Transpositions and inversions were detected as shown in Figure 2.6. The final conserved segments left after all inversions and transpositions had been found were assumed to be segments whose breakpoints were due to translocations. The final conserved segments were manually edited where, for example, two segments were close in the *C. elegans* genome and probably were the same conserved segment.

Because the lengths of those transpositions involving *C. elegans* genes whose *C. briggsae* orthologues have not yet been sequenced can be measured only in units of *C. elegans* genes (Figure 2.6 A), the sizes of all inversions and transpositions have been given in terms of the number of *C. elegans* genes. If a transposition had occurred within an inverted segment, the size of the inversion was taken to include



the transposed genes; likewise, if an inversion had occurred within a transposed segment, the size of the transposition was taken to include the inverted genes.

### 2.5.6 Testing Whether Breakpoints Are Associated with Repeats

The positions of 33 dispersed repeat families in the *C. elegans* genome were downloaded from [http://www.sanger.ac.uk/Projects/C\\_elegans/WORMBASE/GFF\\_Files.shtml](http://www.sanger.ac.uk/Projects/C_elegans/WORMBASE/GFF_Files.shtml). The arrangement of these dispersed repeats into compound repeats was taken from [http://www.sanger.ac.uk/Projects/C\\_elegans/repeats/](http://www.sanger.ac.uk/Projects/C_elegans/repeats/). To group *C. elegans* proteins into families, we compared Wormpep to itself using BLAST (Altschul et al., 1997) with the SEG filter (Wootton and Federhen, 1996). Proteins A, B, and C were assumed to belong to the same family if A hit B with an *E*-value of  $< 10^{-100}$  and B hit C with an *E*-value of  $< 10^{-100}$ .

## 2.6 ACKNOWLEDGEMENTS

We thank the Genome Sequencing Center, Washington University School of Medicine, St. Louis for allowing us to use DNA sequence data before publication. This work was supported by Enterprise Ireland and Science Foundation Ireland. Many thanks to Karsten Hokamp, Simon Wong, and Cathal Seoighe for useful discussions and advice. A special thanks to Aoife McLysaght for help with the Takezaki method, and to Andrew Lloyd, Noel O'Boyle, Richard Durbin, and three anonymous reviewers for critical reading of the manuscript.



## Chapter 3

# The *Caenorhabditis briggsae* Gene Set

During 2001 the Washington University Genome Sequencing Center and the Sanger Institute sequenced the *Caenorhabditis briggsae* genome, producing a high quality draft that covers 98% of the ~104-Mb genome. I was fortunate to collaborate on the *C. briggsae* Sequencing Project during my Ph.D.. This chapter consists of sections I wrote for the *C. briggsae* genome paper describing my work on the project (Stein et al., 2003; to be published in the November 2003 issue of *PLoS Biology*).

### ABSTRACT

We predict about 19,500 protein coding genes in the *C. briggsae* genome, roughly the same number of genes as in *C. elegans*. Of the *C. briggsae* genes, 12,100 have clear *C. elegans* orthologues, a further 6500 have one or more clearly detectable *C. elegans* homologues, and about 800 genes have no detectable matches in *C. elegans*. Among the introns in orthologue pairs, 6579 (9%) are species-specific introns, two-thirds of which are *C. elegans*-specific. The *C. briggsae* draft sequence will greatly improve the annotation of the *C. elegans* genome. Based on similarity to *C. briggsae*, we found strong evidence for 1300 new *C. elegans* genes.

### 3.1 INTRODUCTION

The compactness of the 100-Mb *C. elegans* genome facilitates *ab initio* gene prediction methods, but even the best of these fails to find some genes, and boundaries of genes and exons remain problematic (Reboul et al., 2003). As many as 50% of *C. elegans* gene predictions contain major or minor errors (Reboul et al., 2003). One motivating factor for sequencing the entire *C. briggsae* genome was the promise that comparison between the two genomes would help to correct *C. elegans* predicted gene structures. We



describe here how we predicted the *C. briggsae* gene set, and compared conserved regions in *C. briggsae* and *C. elegans* genes to identify potential errors in *C. elegans* predictions.

## 3.2 RESULTS

### 3.2.1 Protein Coding Genes

*In collaboration with Laura Clarke*<sup>1</sup>, *Michael Brent*<sup>2</sup>, *Chaochun Wei*<sup>2</sup>, *LaDeana Hillier*<sup>3</sup>, and *Todd Harris*<sup>4</sup>

Different programs for predicting protein coding genes agree well when predicting exons, but often disagree on the grouping of exons into genes (Reese et al., 2000). A common procedure for overcoming this problem is to predict genes using several programs, and to compare their output to choose one prediction for each gene (Goff et al., 2002; Rogic et al., 2002). To create the *C. briggsae* gene set, we used the concordance of predictions between *C. elegans* and *C. briggsae* to select the prediction for each gene that was most likely to be correct.

We predicted genes in the *C. briggsae* genome using the programs Genefinder (version 980506; Phil Green, unpublished software), Fgenesh (Salamov and Solovyev, 2000), Twinscan (Korf et al., 2001), and the Ensembl annotation pipeline (Clamp et al., 2003). These programs use a variety of gene prediction methods, including *ab initio* predictions (Genefinder, Fgenesh), EST- and protein-based comparisons (Ensembl), and sequence conservation metrics (Twinscan). We predicted genes in the *C. elegans* genome by combining hand-curated gene structures from WormBase release WS77 (Stein et al., 2001) with *ab initio* predictions from Genefinder and Fgenesh. For technical reasons, we were unable to run Twinscan on the *C. elegans* genome, while the Ensembl method of predicting genes based on matches to previously predicted proteins meant that an Ensembl *C. elegans* gene set would be a duplicate of the hand-curated WormBase set.

The four gene prediction programs agreed well on the position of *C. briggsae* exons (80% of exons predicted identically by two or more programs; 26% predicted identically by all four programs), but disagreed on whole-gene predictions (38% of genes predicted identically by two or more programs, just 4% predicted identically by all four programs). A similar pattern was seen for the genes predicted in *C. elegans*.

To select from overlapping predictions produced by different programs, we reasoned that the gene models most likely to be correct are those that maximise the similarity between predictions in *C. briggsae* and *C. elegans* (Figure 3.1). For each *C. briggsae* gene that had multiple overlapping but inconsistent

---

<sup>1</sup>Wellcome Trust Sanger Institute, Hinxton, United Kingdom

<sup>2</sup>Department of Computer Science and Engineering, Washington University at St. Louis, St. Louis, Missouri, USA

<sup>3</sup>Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri, USA

<sup>4</sup>Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA



predictions, we chose the prediction having the most extensive similarity to the matching *C. elegans* prediction. Likewise, from all the predictions for a *C. elegans* gene, we chose the prediction having the most extensive similarity to its *C. briggsae* match. The extent of similarity was measured by the fraction of the *C. briggsae* prediction that aligned to the matching *C. elegans* prediction at the protein level. We call the gene sets produced by this approach “hybrid” sets, because they consist of a mixture of gene predictions from different programs. Our procedure selected predictions for both species simultaneously, yielding *C. briggsae* and *C. elegans* hybrid gene sets. The gene sets were then filtered to remove transposons and putative pseudogenes.

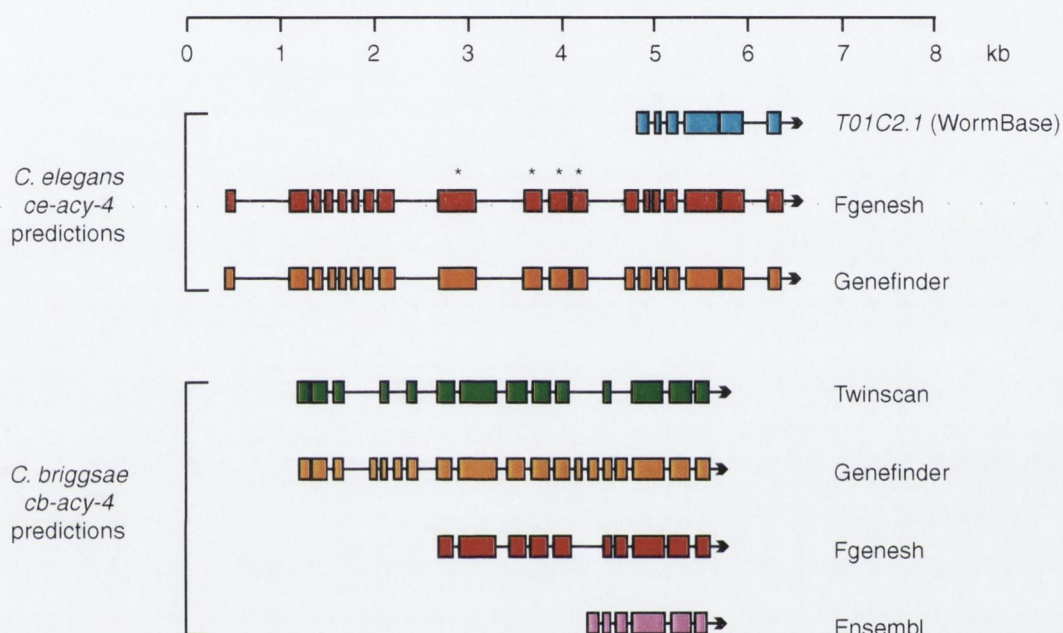


Figure 3.1: Joint refinement of *C. elegans* and *C. briggsae* gene models: *acy-4*. When annotating the *C. briggsae* and *C. elegans* *acy-4* orthologues, we chose the Genefinder *ce-acy-4* prediction and the Genefinder *cb-acy-4* prediction, because out of the 12 possible combinations of a *C. briggsae* and a *C. elegans* prediction, this pair show the most similarity to each other. Coding sequence conservation between *cb-acy-4* and *ce-acy-4* provides evidence for as many as 12 additional N-terminal exons in the Genefinder *ce-acy-4* prediction compared to *T01C2.1*, the WS77 WormBase *ce-acy-4* prediction. Subsequently, four of the additional N-terminal exons that were predicted by Fgenesh and Genefinder were confirmed by new EST data (marked with asterisks).

To assess the accuracy of the gene prediction programs in *C. elegans*, we made a “gold standard” set of *C. elegans* gene predictions: 2257 genes from WormBase WS77 for which every base and intron-exon junction has been confirmed by cDNA or EST data. Genefinder made 2309 predictions that overlapped a gold standard gene, of which 1280 (53%) contained all confirmed bases and introns. Fgenesh made 2742 predictions that overlapped a gold standard gene, of which 1230 (45%) contained all the confirmed data. We also used the gold standard to assess our selection procedure. For *C. elegans* genes in the gold standard set, the selection procedure chose the correct gene model for 92% of gold standard genes, choosing an alternative (incorrect Fgenesh or Genefinder) model 8% of the time. We could not assess



the accuracy of the gene prediction programs or selection procedure in *C. briggsae*, because we lack an independent data set to create a gold standard.

The final transposon-and-pseudogene-filtered *C. briggsae* gene set contains 19,507 genes, and the transposon-and-pseudogene-filtered hybrid *C. elegans* gene set contains 20,621 genes. Some of the gene predictions taken from WormBase WS77 have alternative splices, so the 20,621 *C. elegans* genes have 21,578 different splice variants. There is little EST data for *C. briggsae*, so we are currently unable to predict alternative splices in *C. briggsae*.

In order to compare the transposon-and-pseudogene-filtered *C. briggsae* and *C. elegans* hybrid gene sets to the *C. elegans* WS77 gene set, we applied our transposon and pseudogene filtering step to the *C. elegans* WS77 gene set. This removed 619 genes to create a “pruned” WS77 set of 18,808 genes and 19,791 splices. This pruned set is henceforth called WS77\*. Some of the predictions discarded by our filtering step may include real exons, since 29 (9%) of the 316 putative pseudogenes in *C. elegans* WS77 that were discarded have been partially or fully confirmed by EST or cDNA data.

Data files containing the *C. briggsae* sequence and gene predictions can be found at <ftp://ftp.wormbase.org/pub/wormbase/briggsae/>. The results can also be browsed at <http://www.wormbase.org/>.

### 3.2.2 Comparing the *C. briggsae* and *C. elegans* Gene Sets

The *C. briggsae* gene set (19,507 genes), the *C. elegans* WS77\* gene set (18,808 genes) and the *C. elegans* hybrid gene set (20,621 genes) all contain about the same number of genes. The recent WormBase *C. elegans* release WS103 (June 2003; ~19,600 curated genes) also has a similar number.

The unspliced lengths of genes are roughly the same in the two species (*C. briggsae* median 1.9 kb, *C. elegans* WS77\* 1.9 kb; Table 3.1). The total length of the *C. briggsae* genome occupied by the 19,507 genes, including their introns, is 56 Mb (54% of the 102 Mb assembly) — about the same fraction of the *C. elegans* genome occupied by the WS77\* gene set. Thus the larger size of the *C. briggsae* genome (by ~4 Mb) is not due to an increase in the number or size of protein coding genes (but rather to repetitive DNA; Stein et al., 2003).

The *C. elegans* gene sets have slightly more introns than the *C. briggsae* hybrid set (Table 3.1). Some extra introns may be due to hand-curation of the WS77 gene set, since extra exons that were missed by gene prediction software are added during curation. However, as shown in *C. briggsae*-*C. elegans* *Orthologues* (below), a portion of the intron differences are true evolutionary changes.



	<i>C. briggsae</i>	<i>C. elegans</i> WS77*	<i>C. elegans</i> hybrid
Number of genes	19,507	18,808	20,621
Median gene length	1.90 kb	1.91 kb	1.83 kb
Summed length of genes	55.7 Mb	52.5 Mb	55.6 Mb
Average gene density	5.4 kb per gene	5.3 kb per gene	4.9 kb per gene
Number of exons	114,339	118,045	125,702
Median exon size	150 bp	150 bp	150 bp
Median exons per gene	5	5	5
Median coding length/gene	0.98 kb	1.03 kb	1.00 kb
Summed length of exons	24.1 Mb	24.4 Mb	25.6 Mb
Number of introns	94,832	99,237	105,081
Median intron size	54 bp	66 bp	67 bp
Median intron length/gene	0.75 kb	0.76 kb	0.74 kb
Summed length of introns	31.6 Mb	28.1 Mb	30.0 Mb

Table 3.1: Comparison of the *C. briggsae* and *C. elegans* protein coding gene sets.

### 3.2.3 *C. briggsae*-*C. elegans* Orthologues

*In collaboration with Todd Harris*<sup>1</sup>

We searched for orthologues between the 19,507 *C. briggsae* genes and the 18,808 *C. elegans* WS77\* genes. A gene in one species can have multiple orthologues in another species if the gene has duplicated since the species diverged. However, we used the simpler definition of a pair of genes that have a common ancestor and are in a one-to-one correspondence between two species.

We found orthologues by searching for *C. briggsae*-*C. elegans* gene pairs that were each other's top BLASTP (Altschul et al., 1997) match in the opposite species. We identified 11,255 such gene pairs. We then used conserved gene order between the two species: we identified non-reciprocal-best matches that were flanked by orthologous genes. This step netted 900 more orthologues. The final set of 12,155 orthologues includes 62% of *C. briggsae* genes and 65% of *C. elegans* WS77\* genes.

The median percent identity between orthologous proteins is 80%, similar to the level of divergence between human-mouse orthologues (median 79%; Waterston et al., 2002). The *C. briggsae*-*C. elegans* orthologues are very similar in terms of exon length (median 0.15 kb in both species), coding length per gene (median 1.14 kb in *C. elegans* vs. 1.11 kb in *C. briggsae*), and gene length (median 2.29 kb in *C. elegans* vs. 2.19 kb in *C. briggsae*). Orthologues are longer than the overall set of predicted genes (median 1.90 kb in *C. elegans*), which suggests that the nonorthologous gene set includes some truncated or split gene predictions.

<sup>1</sup>Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA



We searched for *C. elegans* genes that contain introns that are absent from their *C. briggsae* orthologues, and *vice versa*. To do this, we aligned orthologous proteins, and searched for cases where a single exon in one species aligns to two adjacent exons in the other species. We found 6579 species-specific introns among the 60,775 introns in the orthologue pairs: 4379 *C. elegans*-specific introns and 2200 *C. briggsae*-specific introns. This ~2-fold ratio agrees with that reported by Kent and Zahler (2000) using a smaller data set.

### 3.2.4 Estimating the *C. briggsae*-*C. elegans* Divergence Date

Using the divergence of the nematodes from the arthropods at 800–1000 million years ago (Mya; Blaxter, 1998; Brooke, 1999) to calibrate the molecular clock, we estimated the *C. briggsae*-*C. elegans* divergence date from 338 sets of orthologues. Each set comprised a *C. elegans* gene, and its one-to-one orthologues from *C. briggsae*, *Anopheles* and human. When the nematode-arthropod divergence is taken to be 800 Mya, a 95% confidence interval for the median *C. briggsae*-*C. elegans* speciation date is 78–90 Mya. If the nematode-arthropod divergence is taken to be 1000 Mya, the interval becomes 97–113 Mya.

Our best estimate of the *C. briggsae*-*C. elegans* speciation date is therefore ~80–110 Mya. This confidence interval is tighter than our previous estimate of 50–120 Mya made using 92 sets of orthologues from the then 13% sequenced *C. briggsae* genome (Coghlan and Wolfe, 2002). The current estimate is probably more accurate due to both a larger sample size, and improved *C. briggsae* gene predictions and orthologue assignments. Interestingly, recent studies date the human-mouse divergence to 65–75 Mya (Waterston et al., 2002), so the *C. briggsae*-*C. elegans* divergence was at the same time or up to ~50 Myr before the rodent-primate divergence.

### 3.2.5 *C. briggsae*-*C. elegans* Paralogues and Orphans

Roughly a third of *C. elegans* and *C. briggsae* proteins could not be assigned orthologues. Among these are 4545 (23%) *C. elegans* WS77\* genes and 5211 (28%) *C. briggsae* genes that have multiple BLASTP matches in the opposite species. These are members of gene families (examined by Jason Stajich in the *Gene Families* section of Stein et al., 2003).

The remaining 2108 (11%) *C. elegans* and 2141 (11%) *C. briggsae* genes do not have any BLASTP hit of  $E$ -value  $< 10^{-10}$  in the opposite genome, and so are candidate species-specific genes, or “orphans.” However, many of these are simply genes that have evolved rapidly. Lowering the BLASTP threshold to  $E$ -value  $< 10^{-5}$  finds 785 *C. briggsae* proteins that have a weak *C. elegans* match. An additional 11 proteins have a strong TBLASTN match to the *C. elegans* genomic sequence; this *C. elegans* match must either be a *C. elegans* gene that is missing from the predicted gene set or a pseudogene. Another 538 *C. briggsae* genes were found by TRIBE (Enright et al., 2002) to belong to rapidly-evolving shared *C. briggsae*-*C. elegans* gene families (see *Gene Families* in Stein et al., 2003).



	WS77	WS103
New genes	1275	985
New exons in existing genes	1763	1243
Exon extensions in existing genes	1115	845
Exon deletions in existing genes	2093	1600
Exon truncations in existing genes	1675	1114

Table 3.2: Updating the *C. elegans* gene set using *C. briggsae* similarity. We have catalogued possible improvements to *C. elegans* gene models, for both the WS77 gene set used by Stein et al. (2003) in most analyses, and also for the more recent WS103 gene set (June 2003). For WS103, we only have catalogued possible changes for the 15,943 *C. elegans* WS103 gene models that did not change between WS77 and WS103.

This leaves 807 *C. briggsae* proteins that do not have any BLASTP match in the opposite species (of  $E$ -value  $< 10^{-5}$ ), and that do not belong to a shared *C. briggsae*-*C. elegans* gene family. Similarly, we found 1061 *C. elegans* orphans. Of these, 695 *C. briggsae* genes and 963 *C. elegans* genes have at least two exons, and so are less likely than are single-exon predictions to be pseudogenes or mispredictions. Of the *C. elegans* orphans, the gene structures of 208 (22%) orphans have been partially or fully confirmed by EST or cDNA data.

### 3.2.6 Using *C. briggsae* Sequence to Improve *C. elegans* Annotation

*In collaboration with LaDeana Hillier<sup>1</sup> and John Speith<sup>1</sup>*

The *C. elegans* genome now totals 100,273,501 bases (WS103 release; June 2003) and consists of six contiguous segments of DNA corresponding to the six *C. elegans* chromosomes. The last gap in the sequence was closed in November 2002. Since the publication of the *C. elegans* genome (The *C. elegans* Sequencing Consortium, 1998), the gene set has been extensively hand-curated. Between the WS17 WormBase release in April 1999 and the WS77 release in April 2002 (analysed by Stein et al., 2003), WormBase curators made manual changes to ~6300 genes (D. Lawson, pers. comm.).

To investigate the potential of *C. briggsae*-*C. elegans* comparisons for improving the *C. elegans* gene annotations, we compared the *C. elegans* hybrid gene set of 20,621 genes (derived from our comparison of the two species) to the set of 18,808 WS77\* protein coding genes derived from WormBase. The majority (14,011) of the hybrid gene set predictions overlapped perfectly with WS77\* gene predictions. Of course, many of these hybrid predictions were taken directly from WS77. By examining the remaining hybrid predictions, we found strong evidence for 1275 new *C. elegans* genes; 1763 new exons in 1100 existing genes; 2093 exon deletions in 1583 existing genes; 1675 exon truncations in 1502 existing genes; and 1115 exon extensions in 1008 existing genes (Table 3.2).

<sup>1</sup>Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri, USA



Most of the corrections suggested for the WS77 gene set using *C. briggsae* similarity are still applicable to WS103, even after the manual correction of ~3800 *C. elegans* genes between the WS77 (April 2002) and WS103 (June 2003) WormBase releases, prompted in part by the ORF sequence tag (OST) data set of Reboul et al. (2003). Only 290 of the 1275 proposed new hybrid set genes overlap new WormBase gene predictions made since WS77, and 4802 of the 6646 proposed exon changes are in gene structures that have not been edited between WS77 and WS103 (Table 3.2).

We subjected several areas of colinearity to careful hand-curation. In one area containing 33 *C. elegans* predicted genes, the syntenic *C. briggsae* region has 38 predicted genes (Figure 3.2). Rearrangements have broken the syntenic region into three conserved segments, within which gene order and orientation are largely conserved, except for one single-gene inversion (*ZK632.9*). There are 30 one-to-one orthologues in the syntenic block and two one-*C. elegans*-to-two-*C. briggsae* orthologues (*T05G5.6* vs. *CBG10003* and *CBG09979*, and *T05G5.8* vs. *CBG10002* and *CBG09978*), where the two *C. briggsae* orthologues seem to have been duplicated as a block since speciation. The remaining *C. elegans* gene (*CEG09285*) belongs to a *C. elegans*-specific gene family; its closest *C. elegans* paralogue is *F40F12.3*, a gene of unknown function that is nearby on chromosome III. The remaining four *C. briggsae* genes include two members of a *C. briggsae*-specific gene family (*CBG10004* and *CBG09973*); a gene that has a *C. elegans* orthologue on chromosome X (*CBG09992*); and a gene that has no clear *C. elegans* orthologue but has a match on chromosome X (*CBG09973*).

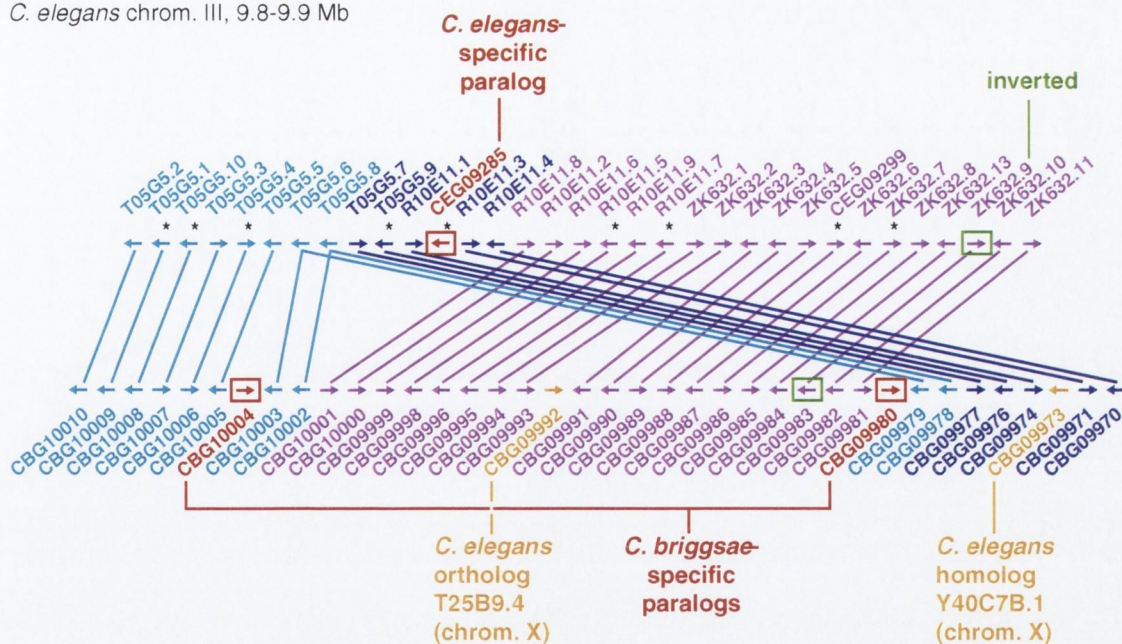
Compared to the *C. elegans* WS77 gene set, the *C. elegans* hybrid set has two extra gene predictions in this region: the *C. elegans*-specific gene *CEG09285*; and *CEG09299*, which is the orthologue of *C. briggsae* gene *CBG09988* (Figure 3.2). The other 31 *C. elegans* genes in this region are in both the *C. elegans* WS77 and hybrid gene sets. However, for 7 of these genes, there are substantial differences between the WS77 and *C. elegans* hybrid gene structures that are supported by *C. briggsae* similarity. These include extra exons (in *T05G5.10* and *T05G5.4*); deletions of WS77 exons (in *T05G5.1*, *T05G5.9* and *ZK632.7*); truncations of WS77 exons (in *R10E11.5*); and extensions of WS77 exons (in *T05G5.1*, *T05G5.4*, and *R10E11.7*). In summary, the analysis of 33 *C. elegans* genes suggested corrections to 7 gene models, proposed 2 missed genes, and confirmed 24 gene models.

### 3.3 DISCUSSION

It is interesting to contrast the *C. briggsae*-*C. elegans* comparison to the recent whole-genome comparison of mouse and human (Waterston et al., 2002). Both pairs of species diverged at about the same time (*C. briggsae* and *C. elegans* 80–110 Mya, human and mouse 75 Mya) and show a similar percent identity between orthologous proteins (80% for *C. briggsae* and *C. elegans*, 79% for human and mouse). However, by multiple measures *C. briggsae* and *C. elegans* are diverging far more rapidly than human and mouse. In human and mouse, 80% of predicted proteins could be assigned to a 1:1 orthologue pair, whereas fewer



*C. elegans* chrom. III, 9.8-9.9 Mb



*C. briggsae* cb25.fpc2234, 0.9-1.0 Mb

Figure 3.2: A region on *C. elegans* chromosome III containing 33 genes, and the syntenic *C. briggsae* region, which has 38 genes. The syntenic region has been broken into three segments, by either one transposition (of the dark blue block) or three overlapping inversions (of a block containing the dark blue and pink genes, of the dark blue block, and of the pink block). Genes that do not have an orthologue in this syntenic region are in orange or red: orthologues are joined by lines. In *C. elegans*, genes that differ substantially in structure between the WS77 and hybrid gene sets are marked with an asterisk. Note that we have not shown the whole of the syntenic region: synteny is conserved further in both the 5' and 3' directions.

than 65% of *C. briggsae* genes could be assigned an orthologue in *C. elegans*. Furthermore, the number of genes lacking a sequence match in the opposite species (orphans) is 4% in *C. briggsae* and *C. elegans*, but less than 1% in human and mouse. Intron gains or losses have occurred at a rate of at least 0.5 per gene since *C. elegans* and *C. briggsae* diverged, while in human and mouse there have been < 0.01 losses or gains per gene in 75 Myr (Roy et al., 2003). Nematodes' faster evolutionary rate may reflect their shorter generation time, which is an order of magnitude less than that of the two mammals.

We estimate that, by using information on *C. briggsae* similarity, the *C. elegans* gene set will be increased by at least ~1300 gene predictions, and that ~2800 exons will be extended or truncated in existing WS77 predictions. The comparative results reported in this paper are currently being used by WormBase curators to improve the *C. elegans* gene set. Especially for poorly-expressed genes, for which EST and mRNA data are not available, and for initial and terminal exons where signals can be difficult to detect, sequence conservation with *C. briggsae* will now provide a primary pointer for *C. elegans* gene structure refinement.



## 3.4 FUTURE WORK

We found ~1650 orphans in the two worms. Some of these orphans may be novel genes that have arisen in one of the two genomes since the species diverged (Long, 2001). However, others of the candidate orphans may not be real orphans at all, but are either pseudogenes that have not yet been deleted, or are very rapidly evolving genes that have diverged so rapidly that the the BLAST and Smith-Waterman algorithms (used in the *Gene Families* section of Stein et al., 2003) cannot recognise their cross-species matches. In *C. elegans*, orphans are clustered on the arms of chromosomes: regions with unusually high rates of chromosomal rearrangement, amino acid substitution, and transposable element insertion (Stein et al., 2003). I am interested in investigating whether the novel worm genes arose as by-products of chromosomal rearrangements, since rearrangements have been implicated in the birth of some novel genes (Long, 2001).

## 3.5 METHODS

### 3.5.1 Protein coding Gene Prediction

We predicted protein coding genes in the *C. briggsae* genome using Genefinder (version 980506; Phil Green, unpublished software), Fgenesh (Salamov and Solovyev, 2000), Twinscan (Korf et al., 2001), and the Ensembl annotation system (Clamp et al., 2003). We also ran Genefinder and Fgenesh on the *C. elegans* genome.

The four gene prediction programs yielded a combined total of 430,575 exon predictions and 73,997 gene predictions in the *C. briggsae* assembly. Many of the predictions from different programs overlapped, so the actual number of exons and genes is far less. The *C. elegans* data consisting of WS77 gene models and Fgenesh and Genefinder predictions totalled 393,529 exon predictions and 61,525 gene predictions.

To select among overlapping predictions produced by different programs, we developed a selection procedure that worked as follows:

1. Many of the exons predicted by different programs overlapped. We took only the longer of any two exons that overlapped by  $\geq 75\%$  of their lengths and were in the same reading frame.
2. We clustered the exons within each species. Two exons were put in the same "exon cluster" if  $\geq 1$  gene prediction program placed them together in a gene prediction. Each exon-cluster consisted of  $\geq 1$  overlapping gene predictions.
3. For each exon-cluster  $X$ , we found the most homologous exon-cluster  $Y$  in the other species. Cluster  $Y$  was the exon-cluster with the top BLASTP (Altschul et al., 1997) hit from any of the exons in  $X$ . For example, for the *C. elegans* exon-cluster containing the *ce-acy-4* gene, its top homologue was the *C. briggsae* exon-cluster containing the *cb-acy-4* gene (Figure 3.1).



4. Each exon-cluster  $X$  consisted of  $n$  overlapping gene predictions  $x_1, x_2, x_3, \dots, x_n$ , where  $n \geq 1$ . We chose one best prediction  $x^*$  for  $X$  in this way:
  - (a) We aligned proteins  $x_1, x_2, x_3, \dots, x_n$  to each of the  $m$  predicted proteins  $y_1, y_2, y_3, \dots, y_m$  in the homologous exon-cluster  $Y$ , using T-COFFEE (Notredame et al., 2000).
  - (b) From each pairwise alignment we calculated a similarity score  $S_{xy} = 0.5(a/L_x + a/L_y)$ , where  $a$  was the number of aligned (not necessarily conserved) amino acids, and  $L_x$  and  $L_y$  the lengths of proteins  $x$  and  $y$ .
  - (c) The best prediction  $x^*$  for  $X$  was that having the highest  $S$  score when aligned to any of  $y_1, y_2, y_3, \dots, y_m$ .
  - (d) If  $X$  was a *C. elegans* exon-cluster, the best prediction  $x^*$  had to agree with experimentally confirmed coding bases and intron-exon junctions in WormBase WS77.

This step produced gene sets for *C. briggsae* and *C. elegans*, which we called the  $G_1$  gene sets.

5. Some exon-clusters did not have a sequence match in the other species. We chose one best prediction for each such exon-cluster by ranking the gene prediction programs, by the fraction of predictions from each program that was selected for the  $G_1$  gene set. The ranking for *C. briggsae* was: Ensembl, Genefinder, Fgenesh, Twinscan. The ranking for *C. elegans* was: the WS77 prediction set, Fgenesh, Genefinder.
6. The predictions chosen were added to the  $G_1$  gene sets, to produce the  $G_2$  gene sets for *C. elegans* and *C. briggsae*.

It is worth noting that there is an unavoidable bias in the way in which our selection procedure produced the  $G_1$  gene sets, which will have affected the ranking of gene prediction programs. Ensembl predicted genes in *C. briggsae* by using similarity to *C. elegans* WS77 genes; therefore, *C. briggsae* Ensembl and *C. elegans* WS77 predictions will tend to have similar structures. Likewise, the *C. briggsae* and *C. elegans* Fgenesh predictions will tend to be similar, because Fgenesh used the same parameters (for example, intron size distribution) to predict both gene sets. Thus, the selection procedure will have selected some *C. briggsae* and *C. elegans* Fgenesh predictions for the  $G_1$  gene sets not because they are more accurate than a *C. briggsae* Twinscan and *C. elegans* Genefinder prediction for that *C. briggsae*-*C. elegans* orthologue pair, but rather because both were predicted by Fgenesh. Therefore, while we used the ranking within our selection procedure, it cannot be used as a comparison of the four gene prediction programs' performance.

The  $G_2$  gene sets were filtered to remove transposons and putative pseudogenes:

1. as described under *Repeat Families* in Stein et al. (2003), we removed transposable element genes;
2. a prediction was taken to be a pseudogene if it was very short or lacked any sequence match:



- (a) if it could only be aligned using T-COFFEE (Notredame et al., 2000) to < 25% of the lengths of its top two matches in *Caenorhabditis* or in SwissProt 40.38 (Boeckmann et al., 2003);
- (b) if it did not have any BLASTP hit in *Caenorhabditis* or SwissProt, of  $E$ -value <  $10^{-10}$  with the SEG filter on (Wootton and Federhen, 1996), or <  $10^{-20}$  with SEG off; or
- (c) if it had a within-species match, but no cross-species match, and was < 50 amino acids long.

This yielded the final ( $G_3$ ) gene sets for *C. elegans* and *C. briggsae*.

### 3.5.2 Finding *C. briggsae*-*C. elegans* Orthologues

We ran NCBI BLASTP (Altschul et al., 1997) with the *C. briggsae* protein set as the query database and the *C. elegans* WS77\* protein set as the target database, and *vice versa*. For *C. elegans* WS77\* genes that have alternative transcripts, we only took the longest splice variant.

We found orthologues in this way:

1. We found *C. briggsae*-*C. elegans* gene pairs that were each other's top BLASTP hits. We required the BLASTP hits to have an  $E$ -value of <  $10^{-10}$  with the SEG filter (Wootton and Federhen, 1996) on, or <  $10^{-20}$  with SEG off. Furthermore, to avoid assigning paralogues to orthologue pairs, the top hit had to have an  $E$ -value  $10^5$  times lower (more significant) than the next best hit.
2. We found additional orthologues by analysing conserved gene order. We found syntenic blocks by looking for orthologues *A* (found in step 1) that were nearby to orthologues *B* (also found in step 1) in both species. We identified *C. briggsae*-*C. elegans* gene pairs within the *A*-*B* syntenic block that were each other's top BLASTP hits within the *A*-*B* block (although not each other's top BLASTP hits within the genome). To avoid assigning paralogues to orthologue pairs, the top hit had to have an  $E$ -value  $10^5$  times lower (more significant) than the next best hit in the *A*-*B* syntenic block.
3. Furthermore, we identified *C. briggsae*-*C. elegans* gene pairs that were each other's top BLASTP hits and that were within 100 kb of orthologues *C* (found in step 1) in both species.

### 3.5.3 Detecting Intron Gain and Loss in Orthologues

We used T-COFFEE (Notredame et al., 2000) to align all *C. briggsae*-*C. elegans* orthologue pairs. We then searched the alignments for cases where exon *i* in species *A* aligned well to two adjacent exons *j* and *k* in species *B*. To ensure that orthologous exons were matched properly, we required that exons *i* and *j*, and exons *i* and *k*, had to consist of identical or conserved amino acids across at least 20% of the shorter exon.



### 3.5.4 Estimating the *C. briggsae*-*C. elegans* Divergence Date

We downloaded human and *Anopheles gambiae* proteins from <http://www.ensembl.org/> in December 2002 (human release 9.30 and mosquito release 9.1; Hubbard et al., 2002). We took the longest alternative splice for each of the 22,980 human genes and 15,088 *Anopheles* genes. To identify *C. elegans*-human orthologues, we compared the *C. elegans* WS77\* protein set to the human proteins using BLASTP (Altschul et al., 1997) with the SEG filter (Wootton and Federhen, 1996). A *C. elegans* gene and human gene were considered one-to-one-orthologues if they were each other's top BLASTP hits, and hit each other with *E*-values of  $< 10^{-20}$ , where the second best hit in each species had to have an *E*-value a factor of  $> 10^{20}$  greater (less significant) than the best hit. In this way we identified 1914 *C. elegans*-human and 2498 *C. elegans*-*Anopheles* orthologues, while 11,255 *C. briggsae*-*C. elegans* one-to-one orthologues were found by identifying mutual-best BLASTP hits as described above. For 1397 *C. elegans* genes, we had a *C. briggsae*, a human and a mosquito orthologue. For each of the 1397 quartets, we aligned the four proteins using CLUSTALW (Thompson et al., 1994), and made a guide-tree using protdist and neighbor from the PHYLIP package (Felsenstein, 1993). For each orthologue set, the alignment and guide-tree were used as input for Gu and Zhang's (1997) program GAMMA, which estimated an  $\alpha$  parameter for the  $\Gamma$  distribution used to correct for rate variation among amino acid sites. For 148 trees, GAMMA could not estimate the  $\alpha$  parameter. For the remaining 1249 trees, we used the two-cluster test (Takezaki et al., 1995) to check for unequal rates between lineages, taking human to be the outgroup to *Anopheles* and *Caenorhabditis* (Aguinaldo et al., 1997); 338 trees passed the test at the 5% significance level. For each of these 338 trees, the branch lengths were re-estimated under the assumption of rate constancy, using Takezaki and Nei's (1995) program with the  $\Gamma$  correction for multiple hits. We calibrated the linearised trees by taking the nematode-arthropod divergence to be 800–1000 Mya (Blaxter, 1998; Brooke, 1999).

### 3.5.5 Using *C. briggsae* Sequence to Improve *C. elegans* Annotation

We used TBLASTN searches (Altschul et al., 1997) of the *C. briggsae* genome to identify gene model changes suggested by the *C. elegans* hybrid gene set, that are absent in the WS77 gene set and that are strongly supported by *C. briggsae* similarity. We only considered new hybrid exons, and extensions, truncations or deletions of existing WS77 exons, where the new/deleted region was  $\geq 5$  amino acids long. We considered there to be strong evidence for a change if:

- a new hybrid exon (absent from the WS77 gene set) had a TBLASTN hit of *E*-value  $< 10^{-3}$  in the *C. briggsae* genome, which covered  $\geq 10$  amino acids of the new exon;
- an extended hybrid exon (present but shorter in the WS77 gene set) had a TBLASTN hit of  $< 10^{-3}$  in the *C. briggsae* genome, which covered  $\geq 10$  amino acids of the extended part;
- a deleted exon (present in the WS77 gene set but not the hybrid gene set) did not have any TBLASTN hit of *E*-value  $< 0.1$  covering  $\geq 5$  amino acids of the WS77 exon;



- a truncated hybrid exon (present but longer in the WS77 gene set) did not have any TBLASTN hit of  $E$ -value  $< 0.1$  covering  $\geq 5$  amino acids of the truncated part.

### 3.6 ACKNOWLEDGEMENTS

This work was supported by Science Foundation Ireland.

Many people collaborated on this work: Laura Clarke<sup>1</sup>, who predicted *C. briggsae* genes using the Ensembl pipeline; Michael Brent<sup>2</sup> and Chaochun Wei<sup>2</sup>, who predicted *C. briggsae* genes using Twinscan; LaDeana Hillier<sup>3</sup>, who predicted *C. briggsae* and *C. elegans* genes using Genefinder, and who along with John Speith<sup>3</sup>, manually inspected a sample of the *C. elegans* hybrid predictions; and Todd Harris<sup>4</sup>, who predicted *C. briggsae* and *C. elegans* genes using Fgenesh, and with whom I worked on finding *C. briggsae*-*C. elegans* orthologues.

A very special thanks to Richard Durbin<sup>1</sup>, who allowed me to join this project and work at the Sanger Institute for two months; to my supervisor Ken Wolfe, who kindly funded my two-month stay in Cambridge; and to Lincoln Stein<sup>4</sup>, who edited and organised the paper (Stein et al., 2003).

---

<sup>1</sup>Wellcome Trust Sanger Institute, Hinxton, United Kingdom

<sup>2</sup>Department of Computer Science and Engineering, Washington University at St. Louis, St. Louis, Missouri, USA

<sup>3</sup>Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri, USA

<sup>4</sup>Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA



## Chapter 4

# Origins of Novel Introns in *Caenorhabditis*

### ABSTRACT

The genomes of the nematodes *Caenorhabditis elegans* and *C. briggsae* both contain about 100,000 introns, of which about 6000 are unique to one species. To study the origins of new introns, we used a method involving phylogenetic comparisons to animal orthologues and other nematode paralogues to identify cases where an intron content difference between *C. elegans* and *C. briggsae* was caused by intron insertion rather than deletion. We identified 86 recently gained introns in *C. elegans* and 42 in *C. briggsae*. Novel introns have a stronger exon splice site consensus sequence than the general population of introns, and they show the same preference for phase 0 sites in codons over phases 1 and 2 as seen in the general population. More of the novel introns are inserted in genes that are expressed in the *C. elegans* germline (61% of genes into which novel introns have inserted) than expected by chance (42% of all genes;  $P = 0.003$ ). As compared to a control set of introns, the novel introns in *C. briggsae* are more likely to contain a repeat element (1.9-fold;  $P = 0.004$ ), and the ends of the intron are more likely to be close to the ends of the repeat element (1.6-fold;  $P = 0.04$ ). Similar but weaker trends are seen in *C. elegans*. Our results narrow down the probable mechanism of intron gain to just two of the five hypothesised mechanisms: transposon insertion and reverse-splicing of a pre-existing intron. We propose an experiment to distinguish between these two hypotheses.

### 4.1 INTRODUCTION

How introns spread within and among genes remain central but largely unresolved questions in evolutionary biology (Gilbert, 1978; Logsdon et al., 1998; Logsdon, 1998). Few proven cases of recent intron invasion are known (Logsdon et al., 1998). However, there is compelling evidence for intron gain. Logsdon et al. (1995) sequenced the triose-phosphate isomerase gene from many animals and found in some



cases, an intron position in one species was not shared with any other species, such that its phylogenetic distribution could be explained either by a single insertion or up to 12 losses. Other convincing examples of recent intron gain have been found in the *SRY* gene of dasyurid marsupials (O'Neill et al., 1998); in the fruitfly xanthine dehydrogenase gene (Tarrío et al., 1998); in the globin genes of midges (Hankeln et al., 1997); in the rice catalase gene (Frugoli et al., 1998); and in chemoreceptor genes of the nematode *Caenorhabditis elegans* (Robertson, 2001).

Logsdon et al. (1995) noticed that nematode genes have a particularly high rate of intron gain compared to other animals. By comparing the whole *C. elegans* genome to 8% of that of its sister species *C. briggsae*, Kent and Zahler (2000) found evidence of ~250 introns present in one species but not in the other. Recently, we reported that in 12,155 orthologous gene pairs in the whole genomes of *C. elegans* and *C. briggsae*, there are 4379 *C. elegans*-specific introns and 2200 *C. briggsae*-specific introns (Stein et al., 2003). We estimated that intron gains or losses have occurred at a rate of at least 0.005 per gene per Myr since *C. elegans* and *C. briggsae* diverged, which far exceeds the rate in chordates (Stein et al., 2003). Intron-exon structure seems to be in flux across the entire Phylum Nematoda: in 11 orthologous genes compared between *C. elegans* and its distant relative *Brugia malayi*, only 50% of *C. elegans* introns are conserved in *B. malayi*, and 25% of *B. malayi* introns conserved in *C. elegans* (Guiliano et al., 2002).

Despite strong evidence that intron gains occur, the mechanism is unknown. Here, we searched for novel introns that have been gained after the divergence of *C. elegans* and *C. briggsae*. Our results narrow down the probable mechanism of intron gain to just two of the five hypothesised mechanisms: transposon insertion (Crick, 1979; Cavalier-Smith, 1985) and reverse-splicing of a pre-existing intron (Sharp, 1985). Finally, we propose an experiment to distinguish between these two hypotheses.

## 4.2 RESULTS

### 4.2.1 Identification of Novel Introns

We considered a *C. elegans* or *C. briggsae* intron to be novel if it is absent from the gene's orthologues in *C. briggsae*, *C. elegans*, the nematode *Brugia malayi*, chordates (man and mouse), and arthropods (fruitfly and mosquito). To ensure that a putative novel intron was almost certainly caused by intron insertion rather than by deletion, we drew phylogenetic trees of the gene and its animal and nematode orthologues. We required that there be  $\geq 3$  nodes between the gene and the outgroup (Figure 4.1). Because  $\geq 3$  independent intron losses or one gain could explain the intron distribution, it is more parsimonious to infer intron gain. Furthermore, to ensure that a putative novel intron is very unlikely to be due to intron sliding (Rogozin et al., 2000), the novel intron had to be  $> 5$  amino acids from the nearest intron in any homologue. Using this rigorous approach we found 42 novel introns in *C. briggsae* and 86 in *C. elegans*. The phylogenetic trees, and protein alignments showing the positions of novel introns, can be viewed at



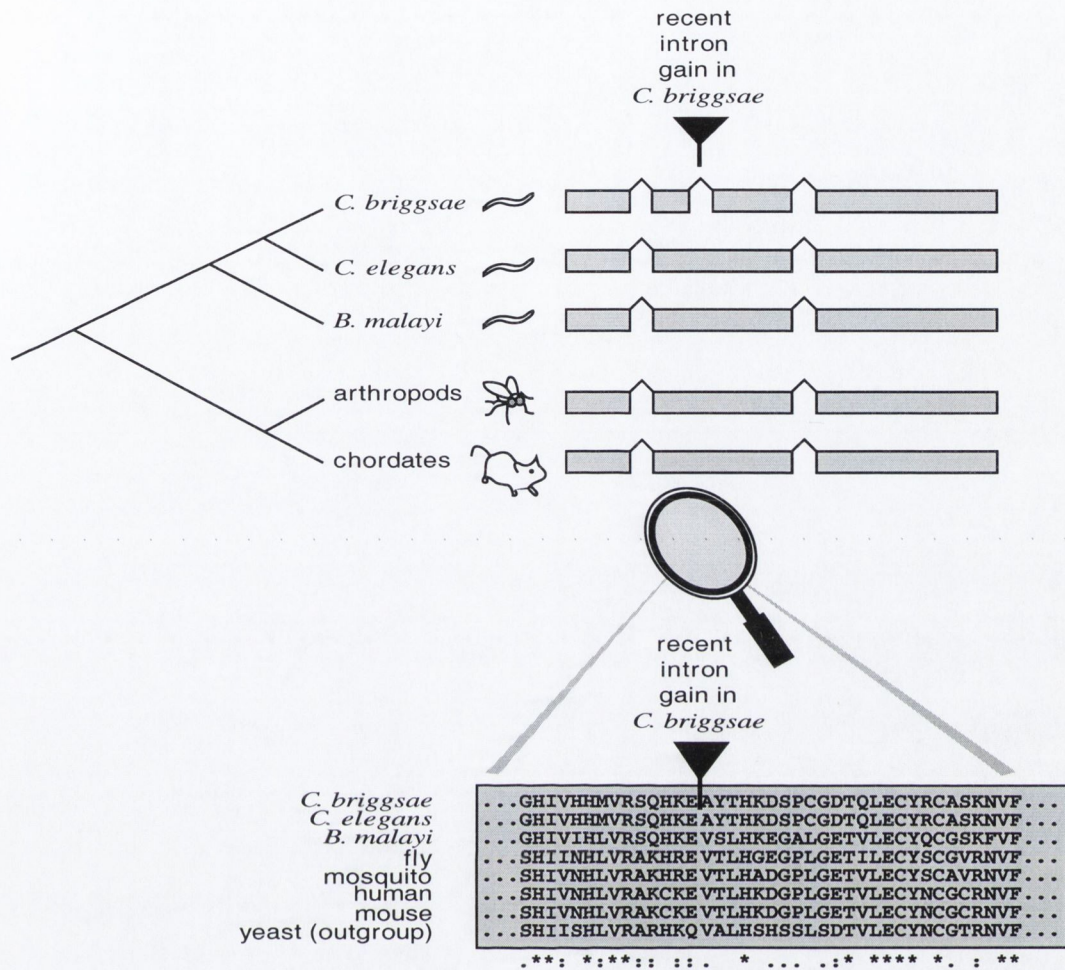


Figure 4.1: Identifying novel introns. To ensure that a putative novel intron was almost certainly caused by insertion rather than by deletion, we drew phylogenetic trees of the gene and its animal and nematode orthologues. We required that there be  $\geq 3$  nodes between the gene and the outgroup. We also required that, in a protein alignment of the gene and its orthologues,  $\geq 5/10$  positions on either side of the intron must be identical or well conserved.

#### 4.2.2 Exon Splice Site Consensus of Novel Introns

To compare the novel introns to the entire *C. elegans* and *C. briggsae* intron populations, we created a control set of introns (see METHODS). In the control set of introns from *C. briggsae*, 55% of introns have A as the second-last exonic base before the 5' splice site, and 59% of introns have G as the last exonic base before the 5' splice site (Figure 4.2). The AG 5' exonic consensus is stronger in the 42 *C. briggsae* novel introns. That is, 71% of novel introns have A as the second-last exonic base (one-sided Fisher's exact test;  $P = 0.03$ ), and 86% have G as the last exonic base (one-sided Fisher's exact test;  $P = 0.0002$ ). This trend is also seen in *C. elegans*: A is the second-last exonic base in 78% of novel but only 56% of control introns (one-sided Fisher's exact test;  $P = 10^{-5}$ ), and G the last exonic base in 84% of novel



compared to 60% of control introns (one-sided Fisher's exact test;  $P = 10^{-6}$ ).

	5'		3'	
	-2 -1		+1 +2	
	... A G		G T ...	
<i>C.e.</i> control	56	60	30	38
<i>C.e.</i> novel	78	84	53	44
<i>C.b.</i> control	55	59	30	37
<i>C.b.</i> novel	71	86	52	40

Figure 4.2: The exon splice site consensus of novel introns in *C. elegans* and *C. briggsae*, compared to the splice site consensus for a control set of introns from the two genomes.

Likewise, for the 3' splice site a higher proportion of *C. briggsae* novel introns have G as the first exonic base after the 3' splice site (52%) compared to control introns (30%; one-sided Fisher's exact test;  $P = 0.002$ ). *C. elegans* novel introns also tend to have G as the first exonic base: 53% of novel introns have G compared to only 30% of control introns (one-sided Fisher's test;  $P = 10^{-6}$ ). Furthermore, the second base in the exon after the 3' splice site tends to be T more often in novel introns than in control introns, but this is not statistically significant (Figure 4.2).

#### 4.2.3 Phases of Novel Introns

An intron has a "phase" of 0 if it is between two codons in a gene, while its phase is 1 if it is after the first base of a codon, or 2 if it is after the second base of a codon. If introns inserted into random positions in genes, novel introns would have an equal probability of having phase 0, 1, or 2. However, of the 42 novel introns in *C. briggsae*, 23 (55%) are phase 0, 12 (29%) phase 1, and 7 (17%) phase 2. That is, the novel introns deviate significantly from having equal proportions of each phase ( $\chi^2$  test;  $P = 0.008$ ). This trend is also seen in *C. elegans*, where 53% of the novel introns are phase 0, 24% phase 1, and 22% phase 2 ( $\chi^2$  test;  $P = 0.0004$ ).

Of the *C. briggsae* control introns, 51% have phase 0, 24% phase 1, and 25% phase 2. Similarly, of the *C. elegans* control introns, 53% have phase 0, 24% phase 1, and 22% phase 2. The distribution of phases of novel introns in *C. briggsae* is not significantly different from that of the control introns ( $\chi^2$  test;  $P = 0.4$ ), and the same is true for novel *C. elegans* introns ( $\chi^2$  test;  $P = 0.8$ ).

#### 4.2.4 Germline Expression of Genes that have Gained Introns

To become fixed, an intron gain must occur in a germline cell or a cell that is going to become one. We investigated whether intron gain also requires expression in the germline. Hill et al. (2000) used oligonucleotide arrays to identify 5951 *C. elegans* genes that are always or sometimes expressed in oocytes.



Of the 78 genes that have gained introns in *C. elegans*, 61 were studied by Hill et al. (2000), while their data set covers 4752 of the genes containing control introns. The proportion of the 61 genes that gained introns that are always/sometimes oocyte-expressed (61%) is significantly greater than the proportion of the 4752 control genes that are always/sometimes oocyte-expressed (42%; one-sided Fisher's exact test;  $P = 0.003$ ). Thus, genes that are expressed in the germline are more susceptible to gaining introns than genes not expressed in the germline.

#### 4.2.5 Repeat Elements in Novel Introns

We tested the hypothesis that novel introns originate from transposable elements (Crick, 1979; Cavalier-Smith, 1985), by testing whether novel introns contain more repeat elements than do control introns. A higher proportion of the 42 *C. briggsae* novel introns (38%) contain repeat elements than do the 18,516 control introns (20%; one-sided Fisher's exact test;  $P = 0.004$ ; Figure 4.3). In *C. elegans*, more of the 86 novel introns (23%) contain repeat elements than do the 19,942 control introns (16%;  $P = 0.05$ ).

If a novel intron originated by insertion of a transposable element, one would expect that initially the entire intron consisted of transposable element DNA. We found that, for *C. briggsae* introns that contain a repeat element, a higher proportion of novel introns contain a repeat element within 25 bp of the intron start or end (69%) than do control introns (44%; one-sided Fisher's exact test;  $P = 0.04$ ). In *C. elegans* the same trend is seen but it is not statistically significant: among introns that contain a repeat element, the repeat is  $\leq 25$  bp from the intron start or end in 55% of novel introns compared to 40% of control introns (one-sided Fisher's exact test;  $P = 0.1$ ).

There are members of 31 different repeat families in the 86 *C. elegans* novel introns. Of these, 21 are putative nonautonomous DNA transposons and the rest are yet-unclassified. The DNA transposon families belong to several different superfamilies: MITEs, HAT, MUDR and mariner. There are members of 36 different repeat families in the 42 novel *C. briggsae* introns, of which 3 are DNA transposons, 1 is a retroelement, and the rest are unclassified. Many of the unclassified repeat families are probably species-specific transposable element families (Stein et al., 2003).

### 4.3 DISCUSSION

#### 4.3.1 Method for Identifying Recently Gained Introns

Logsdon et al. (1998) emphasise that for an intron gain to be convincing, there must be (i) good taxon sampling, to distinguish between intron loss and gain; and (ii) the source of the novel intron must be identifiable. To satisfy the first requirement, we used a phylogenetic approach to detecting novel introns, only retaining cases where there were  $\geq 3$  nodes between the gene that has gained an intron and the outgroup (Figure 4.1). To fulfill Logsdon et al.'s second condition, we only included putative



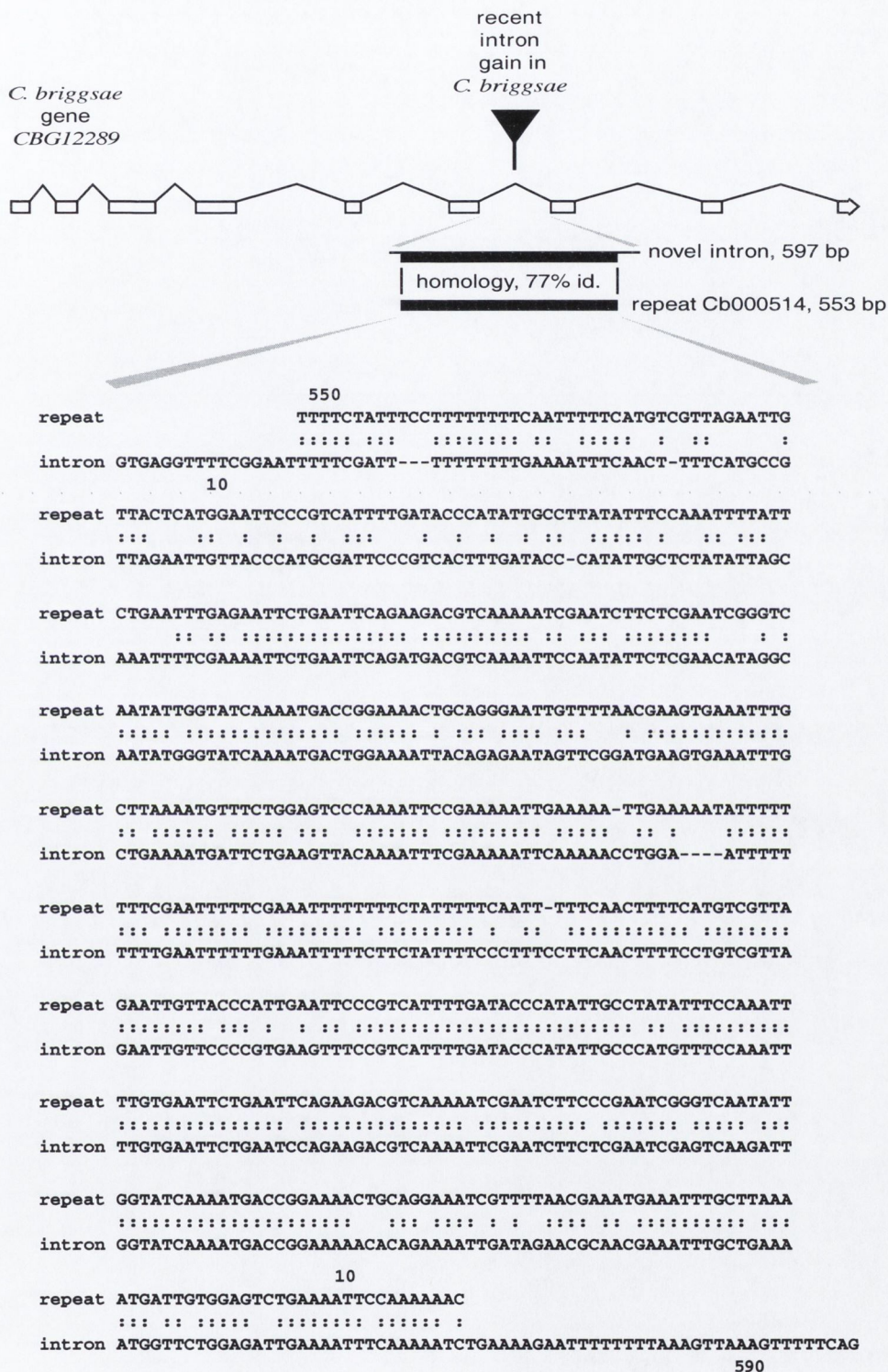


Figure 4.3: An example of a novel intron containing a repeat element: the fourth intron in *C. briggsae* gene *CBG12289* is a novel intron, which contains a *C. briggsae*-specific repeat element Cb000514.



novel introns that are either *C. elegans*-specific or *C. briggsae*-specific, and so have been gained in the 80–110 million years (Myr) since the two species diverged (Stein et al., 2003). Introns evolve at the neutral rate of nucleotide substitution, which is particularly high in *Caenorhabditis* compared to most animals (Aguinaldo et al., 1997). After 20 Myr a 50-bp novel intron will be only 60–80% identical to its source sequence, since the silent substitution rate is 0.01–0.02 substitutions per site per Myr in *Caenorhabditis* (Cutter and Payseur, 2003). Thus, to increase our chances of finding the “molecular smoking gun” revealing novel introns’ source sequences, as Logsdon et al. (1998) put it, we narrowed our search to species-specific novel introns. Such a scan has only been possible since the recent sequencing of the *C. briggsae* genome (Stein et al., 2003) resulted in the first pair of fully sequenced genomes from species in the same animal genus.

### 4.3.2 Rate of Intron Gain in *C. elegans* vs. *C. briggsae*

We found about twice as many novel introns in *C. elegans* (86) as in *C. briggsae* (42). This ratio agrees with previous findings of about twice as many species-specific introns in *C. elegans* as *C. briggsae* (Kent and Zahler, 2000; Stein et al., 2003), and suggests that this trend is mainly due to a two-fold higher rate of intron gain in *C. elegans* rather than to a higher rate of intron loss in *C. briggsae*.

Intron gains may occur more frequently in *C. elegans* if the mutations causing intron gain occur more frequently in *C. elegans*; for example, if intron gains are caused by transposon insertion and there are more active transposons in *C. elegans*. Alternatively, intron gains may occur at the same rate in the two species, but be fixed two-fold more frequently in *C. elegans*. It is yet unclear whether population-genetic factors affect rates of intron gain in different species (Lynch, 2002). *C. elegans* may have a smaller effective population size than *C. briggsae*, because there seems to be slightly less diversity in *C. elegans* populations than *C. briggsae* populations (Graustein et al., 2002; Jovelin et al., 2003). A smaller effective population size in *C. elegans* would result in a higher rate of random fixation of neutral or mildly deleterious mutations such as intron insertions. It will be interesting to see, when we have more genomic sequence from the cross-fertilising *C. remanei*, whether *C. remanei* has gained introns at an even slower pace than *C. briggsae*.

### 4.3.3 Mechanisms of Intron Gain

Five different mechanisms by which novel introns are gained have been proposed. Shortly after the discovery of introns, Crick (1979) hypothesised that novel introns arise by insertion of a transposon (see also Cavalier-Smith, 1985). There is a large body of evidence showing that transposable elements that have inserted into laboratory strains of animals and plants can be spliced, and that the phenotype of the insertion mutant is often wild-type or near wild-type (Giroux et al., 1994; reviewed in Purugganan, 2002). However, there is no evidence that this also occurs on an evolutionary timescale. Rogers (1989) suggested that novel spliceosomal introns may originate by insertion of a group II intron via reverse self-splicing,



but there is no evidence to support this. Rogers (1989) also put forward an alternative hypothesis: that novel introns arise by tandem duplication of an internal fragment of an exon containing AGGT, and that the resultant cryptic splice sites are then activated. Three novel introns in fish genes align well to nearby exon sequence, so may have arisen by this mechanism (Venkatesh et al., 1999). A fourth hypothesis is that a pre-existing spliceosomal intron is reverse-spliced into a new site in the same or a different mRNA, which is then reverse-transcribed to a cDNA which recombines with the genome (Sharp, 1985). Tarrío et al. (1998) attributed three novel introns in the fly xanthine dehydrogenase gene to this mechanism, but others have not been convinced by their analysis (Logsdon et al., 1998). A fifth hypothesis is that an intron-containing mRNA is reverse-transcribed, and that the cDNA recombines with a homologue in the genome which previously lacked an intron at that site (Hankeln et al., 1997). There is strong evidence that introns were gained in midge globin genes by this mechanism (Hankeln et al., 1997).

Intron gain by group II intron insertion is very unlikely to be responsible for the novel nematode introns, because animal mitochondrial genomes do not contain group II introns (Bonen and Vogel, 2001). Intron gain by gene conversion with a homologous intron-containing gene results in the novel intron being gained at the same position as the source intron (Hankeln et al., 1997). We only included novel introns for which there was no intron at the same position in any close homologue, so the novel introns in our data set probably did not arise by this mechanism either. Thus, in the following discussion, we consider whether the remaining three mechanisms could explain our data: transposon insertion; partial exon duplication; and reverse-splicing of a pre-existing intron.

#### 4.3.4 Germline Expression

We found that 61% of *C. elegans* genes that gained introns are expressed in the germline, compared to 42% of control genes (Section 4.2.4). Some novel introns reported in the literature are in genes expressed in the germline, such as the *SRY* gene (O'Neill et al., 1998). Others are in genes that encode key metabolic enzymes, and so are expressed constitutively in all tissues including the germline, such as triose-phosphate isomerase (Logsdon et al., 1995) and catalase (Frugoli et al., 1998). Logsdon et al. (1998) pointed out that if introns are gained by reverse-splicing of a pre-existing intron, one would expect intron gains to occur mainly in germline-expressed genes. Alternatively, if novel introns arise by transposon insertion, the transposons responsible may have an insertion preference for actively transcribed regions of the genome. Such a preference has been observed for transposons such as the *Drosophila* *P*-element (Timakov et al., 2002). If intron gains occur by partial exon duplication, we can see no reason why there would be a bias for germline-expressed genes. Thus, at this point we discarded partial exon duplication as a possible explanation for intron gain in *Caenorhabditis*.

The "exon theory of genes" proposes that primordial genes contained introns, which were involved in exon shuffling in these earliest genes (Gilbert, 1978; reviewed in Roy, 2003). Many of the results reported



as supporting this hypothesis are based on the assumption that ancient proteins shared by prokaryotes and eukaryotes contain older introns on average than do eukaryote-specific genes (for example, Fedorov et al., 2001). However, 53% of *C. elegans* genes containing a novel intron, but only 11% of control genes, have a BLASTP match to a bacterial protein in SwissProt with  $E$ -value  $< 10^{-10}$  (two-sided Fisher's exact test;  $P < 10^{-16}$ ). Likewise, in *C. briggsae* 58% of genes containing a novel intron have a bacterial homologue, compared to 11% of control genes (two-sided Fisher's exact test;  $P = 10^{-12}$ ). Thus, ancient genes gain introns more frequently than do younger genes. In other words, a larger proportion of the introns in ancient genes are probably young compared to in younger genes, whereas the exon theory proposes that a larger proportion of the introns in ancient genes are ancient compared to younger genes. This seems to be because a large fraction of germline-expressed genes are ancient genes (presumably housekeeping genes such as metabolic enzymes): 17% of germline-expressed genes have bacterial homologues compared to 10% of all *C. elegans* genes (two-sided Fisher's exact test;  $P < 10^{-16}$ ). Fedorov et al. (2003) reported that introns at conserved positions in ancient genes tend to be found at protein module boundaries. They interpreted this as evidence of exon shuffling in primordial genes. While our finding of frequent intron gain in ancient genes does not disprove the exon theory of genes, it raises the important question of whether such introns may have been independently gained in phylogenetically distant taxa (Tarrío et al., 2003).

#### 4.3.5 The Proto-Splice Site

Based on 60 putative novel introns in actin and tubulin genes from a broad range of eukaryotes, Dibb and Newman (1989) hypothesised that novel introns tend to insert at the sequence MAG↓R, where ↓ is the insertion site, M=A/C and R=A/G. They called this the "proto-splice site." We found that novel introns in *Caenorhabditis* tend to insert at AG↓G (Section 4.2.2). This agrees with the results of Kent and Zahler (2000), who found that in *Caenorhabditis* the 5' exon AG and 3' exon G consensus is stronger in species-specific introns than in all introns. If introns are gained by reverse-splicing of a pre-existing intron, the spliceosome may insert the novel intron into AG↓G, since this would be the reverse of its normal role of removing an intron from AG↓G. Alternatively, if novel introns arise by transposon insertion, if the transposon produced a target site duplication containing AGG, the resultant intron would be found at AG↓G (Giroux et al., 1994).

We found that novel introns do not insert in random positions in codons, but tend to insert at phase 0 positions (between codons) more than phase 1 or 2 positions (within codons; Section 4.2.3). This agrees with the results of Rogozin et al. (2003), who found that in eight different eukaryotes, putative novel introns have a greater tendency to be in phase 0 than do ancient conserved introns. If novel introns insert into AG↓G, 51% of insertions will be in phase 0 because of the genetic code (Long et al., 1998; see Logsdon, 1998 for discussion). This is close to the fraction of novel introns in phase 0 that we observed: 55% in *C. briggsae* and 53% in *C. elegans*. Thus, the excess of phase 0 introns among the novel introns is



likely to be a result of the tendency of novel introns to insert at AG↓G sites.

#### 4.3.6 The Molecular Smoking Gun

Compared to a control set of *C. briggsae* introns, the novel introns in *C. briggsae* are 1.9-fold more likely to contain a repeat element ( $P = 0.004$ ; Section 4.2.5). Similarly, in *C. elegans*, novel introns are 1.4-fold more likely than control introns to contain repeat elements ( $P = 0.05$ ; Section 4.2.5). At first glance, this result suggests that, in *C. briggsae* at least, novel introns have originated by insertion of a transposon. However, if introns are gained by reverse-splicing of pre-existing introns, and certain *C. briggsae* transposons tend to insert into open chromatin in transcribed genes, we would see a spurious association between repeat elements and novel introns because both insert into expressed genes. That is, transposons may have inserted into novel introns after their birth. Indeed, when we controlled for germline expression in *C. elegans*, there was no significant difference between the fraction of novel introns in germline-expressed genes that contain repeat elements (18%) and the fraction of introns in germline-expressed control genes that contained repeat elements (12%; one-sided Fisher's exact test;  $P = 0.2$ ). Unfortunately, we could not control for germline expression in *C. briggsae* because we do not have *C. briggsae* expression data.

#### 4.3.7 Conclusion

There are two mechanisms for intron gain that are compatible with our data: transposon insertion (Crick, 1979; Cavalier-Smith, 1985) or reverse-splicing of a pre-existing intron (Sharp, 1985). It is of course possible that both occur to different extents. We discarded three other hypothesised mechanisms as being incompatible with our results: intron gain by partial exon duplication (Rogers, 1989); insertion of a group II intron (Rogers, 1989); and gene conversion with an intron-containing homologue (Hankeln et al., 1997).

Since the acid test of a theory is whether it can predict a new experimental result, let us formulate a suitable experiment. To distinguish whether the major mechanism of intron gain is reverse-splicing or transposon insertion, we need a large data set of novel introns that are even younger than those examined here, because younger introns are more likely to have retained homology to their source sequence. If the *C. remanei* genome is sequenced (as Stein et al., 2003 have proposed), we could find *C. remanei*-specific introns, which would be younger than those studied here because *C. remanei* diverged from *C. briggsae* after *C. elegans* and *C. briggsae* diverged (Jovelin et al., 2003). The reverse-splicing model would be favoured if *C. remanei*-specific introns are homologous to other introns, even outside repeat elements. On the other hand, the transposon insertion model would be favoured if (i) after controlling for germline expression, *C. remanei*-specific introns contain more repeat elements than do control introns; (ii) the repeat elements in *C. remanei*-specific introns are nearer the intron ends than are the repeat elements in control introns.



## 4.4 METHODS

### 4.4.1 Sources of Sequence Data

The *C. elegans* protein set was downloaded from <http://www.sanger.ac.uk/Projects/C.elegans/wormpep/> (Wormpep104; July 2003; 19,588 genes). We took the longest alternative splice for each *C. elegans* gene. The *C. briggsae* protein set, consisting of 19,507 proteins, was created as part of the *C. briggsae* Sequencing Project (Stein et al., 2003). The 32,035 human protein sequences from 23,299 genes in Ensembl human release 15.33.1 (Clamp et al., 2003) were downloaded from <ftp://ftp.ensembl.org/>. We also downloaded Ensembl mouse release 15.30.1 (32,911 proteins; 24,948 genes); Ensembl *Drosophila* release 15.3a.1 (18,282 proteins; 13,525 genes); and Ensembl *Anopheles* release 15.2.1 (16,122 proteins; 14,653 genes). SwissProt 41.15 (July 2003; Boeckmann et al., 2003) was downloaded from [ftp://ftp.ebi.ac.uk/pub/databases/sp\\_tr\\_nrdp/fasta](ftp://ftp.ebi.ac.uk/pub/databases/sp_tr_nrdp/fasta).

### 4.4.2 Finding the Closest Homologues of each Nematode Gene

For each *C. elegans* or *C. briggsae* gene, we found its closest homologues in *C. elegans*, *C. briggsae*, human, mouse, fruitfly, and mosquito by homology searches with BLASTP (Altschul et al., 1997), using the SEG filter (Wootton and Federhen, 1996), an effective database size of 10,000,000 and an *E*-value cutoff of  $\leq 10^{-10}$ . We sorted the homologues of a gene in order of significance (increasing *E*-value), and took the most significant hits. As we were taking the top hits:

- if we found a hit of *E*-value  $>10^{-30}$ , or
- if we found a hit with an *E*-value  $>10^{10}$  higher than the previous hit, or
- if we already had 15 sequences,

then that hit and all less significant hits were discarded. These numbers ( $10^{-30}$ , 15,  $10^{10}$ ) were chosen to maximise the number of homologues found for a gene, while excluding distant homologues. We found homologues for 16,590 *C. elegans* genes and 16,438 *C. briggsae* genes.

### 4.4.3 Detecting Intron Gains from Protein Alignments

The proteins in each of the 33,028 groups of homologues were aligned using CLUSTALW (Thompson et al., 1994). We discarded poor alignments, keeping only alignments where  $\geq 80\%$  of the alignment length does not have a gap, and for which  $\geq 40\%$  of the non-gapped part of the alignment consists of identical or conserved residues. We also discarded alignments that contained  $< 4$  sequences, as this is too small a sample to be convincing evidence of intron gain. To detect recently gained introns in a gene, we calculated the position of the gene's introns with respect to the protein alignment of that gene to the gene's closest homologues. If a *C. briggsae* or *C. elegans* gene *A* has an intron  $A_i$  after its  $i^{\text{th}}$  amino acid (taking the position of a phase 1 or phase 2 intron to be before the amino acid whose codon it splits), and



amino acid  $i$  is at the  $j^{\text{th}}$  position of the alignment, then intron  $A_i$  is at the  $j^{\text{th}}$  position of the alignment. We excluded introns in gene  $A$  that fall in unreliable regions of the protein alignment, considering the position of an intron at the  $j^{\text{th}}$  position of the alignment to be reliable if:

- $\geq 5/10$  of the aligned amino acids from  $j - 9$  to  $j$ , and  $\geq 5/10$  of those from  $j + 1$  to  $j + 10$ , are either identical or conserved, and
- there are no gaps (-) in the alignment between positions  $j - 9$  to  $j + 10$ .

Taking only those introns whose positions are reliable, an intron at the  $j^{\text{th}}$  position of the alignment is considered to have been recently gained in  $A$  if there is no intron in any of the homologues of  $A$  from  $j - 4$  to  $j + 5$ . Because a novel intron had to be  $> 5$  amino acids from an intron in any homologue, it is unlikely that any of the putative novel introns are actually not novel introns but rather are cases of intron sliding (Rogozin et al., 2000). We found 244 putative novel introns in *C. elegans* and 124 in *C. briggsae*.

#### 4.4.4 Checking whether Putative Novel Introns are Present in *Brugia malayi*

We checked whether the 368 putative novel introns are present in a distantly related nematode, the filarial nematode *Brugia malayi*, which diverged from the lineage leading to *Caenorhabditis* about 550 Mya (Vanfleteren et al., 1994). The genome of *B. malayi* is currently being sequenced by The Institute for Genomic Research (TIGR; <http://www.tigr.org/tdb/e2k1/bma1/>).

There are no gene predictions available yet for *B. malayi*, so to check whether a *Caenorhabditis* intron is present in *B. malayi*, we ran TBLASTN (Altschul et al., 1997) using the *Caenorhabditis* protein as query. In general, the top *B. malayi* hit consisted of several closely spaced TBLASTN matches (high-scoring segment pairs; HSPs) in the *B. malayi* genome, corresponding to the exons of the *B. malayi* homologue. If a putative novel intron was at amino acid  $i$  in the *Caenorhabditis* protein, we took the intron to be present in *B. malayi* if the top *B. malayi* TBLASTN hit included two nearby exons (HSPs), the first exon ending at amino acid  $i \pm 5$  (with respect to the *Caenorhabditis* protein), and the following exon starting at amino acid  $i \pm 5$ . Likewise, we took a *Caenorhabditis* intron to be absent in *B. malayi* if the top *B. malayi* TBLASTN hit included a large exon (HSP), where amino acid  $i$  was in the middle of the *B. malayi* exon,  $\geq 5$  residues from either end. Of the 244 putative novel *C. elegans* introns, 112 are absent and 73 present in *B. malayi* (59 were ambiguous, because the *B. malayi* TBLASTN hits were weak, or there was no TBLASTN hit in the current *B. malayi* assembly). Of the 124 putative novel *C. briggsae* introns, 57 are absent and 39 present in *B. malayi* (28 ambiguous).

#### 4.4.5 Phylogenetic Support for Intron Gains

Logsdon et al. (1998) emphasise that for an intron gain in a gene to be convincing, one needs a large sample of closely related homologues that lack the intron. To ensure this was so, we constructed phylogenetic trees of each gene containing a putative novel intron and of that gene's homologues:



1. the outgroup for the tree was a SwissProt yeast, bacterial, plant or animal protein, that was clearly more distant from the other proteins in the tree than they were to each other. We did not examine the intron-exon structure in the outgroup (often unknown); the outgroup was only used to root the tree;
2. the proteins for each tree were aligned using T-COFFEE (Notredame et al., 2000), and we made a guide-tree from the alignment using protdist and neighbor (Felsenstein, 1993);
3. the alignment and guide-tree were used as input in Gu and Zhang's (1997) program GAMMA, which estimates an  $\alpha$  parameter for the  $\Gamma$  distribution used to correct for rate variation among amino acid sites;
4. we redrew neighbour-joining trees using protdist and neighbor with the  $\Gamma$  correction for multiple hits, and bootstrapping the trees using 1000 bootstrap replications in seqboot (Felsenstein, 1993);
5. a phylogenetic tree was only taken as acceptable if there were  $\geq 3$  nodes having bootstrap value  $\geq 70\%$  between the outgroup and the gene containing a putative novel intron.

We found phylogenetic support for 42 *C. briggsae* and 86 *C. elegans* putative novel introns. The phylogenetic trees, and protein alignments showing the positions of novel introns, can be viewed at <http://wolfe.gen.tcd.ie/avrill/introns.html> (password = *quereckoneas44*).

#### 4.4.6 Control Set of Introns

To compare the novel introns to the entire *C. elegans* and *C. briggsae* intron populations, we created a control set of introns. This was necessary because some predicted introns are unreliable: their intron-exon boundaries may be wrong, or the gene prediction containing them may be a false positive. We only included introns in our control set for which:

- in protein alignments,  $\pm 10$  amino acids adjacent to the intron's position are well conserved (as required for novel introns);
- 80% of the protein alignment does not have a gap, and  $\geq 40\%$  of the non-gapped part of the alignment consists of identical or conserved residues (as required for novel introns).

The control set consists of 19,942 *C. elegans* introns (20% of all *C. elegans* introns) and 18,516 *C. briggsae* introns (20%).

#### 4.4.7 Detecting Repeat Elements in Introns

To find repeat elements in the *C. elegans* and *C. briggsae* novel introns and control introns, we used fasta (Pearson and Lipman, 1988) with an *E*-value cutoff of 0.05 and ktup of 6, and searched the repeat libraries for *C. elegans* and *C. briggsae* made by Zhirong Bao and Jack Chen (Stein et al., 2003).



## 4.5 ACKNOWLEDGEMENTS

This work was supported by Science Foundation Ireland. We thank the Institute for Genomic Research (TIGR) for generously allowing use of *B. malayi* DNA sequence data before publication. Furthermore, we are grateful to Dr. Richard Durbin and Dr. Lincoln Stein of the *C. briggsae* Sequencing Project, for allowing use of *C. briggsae* genome sequence before publication. We thank Dr. Andrew Hill for kindly providing files containing *C. elegans* germline expression data (Hill et al., 2000).



# Bibliography

- Aguinaldo, A. M., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff, and J. A. Lake (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387, 489–93.
- Akerib, C. C. and B. J. Meyer (1994). Identification of X chromosome regions in *Caenorhabditis elegans* that contain sex-determination signal elements. *Genetics* 138, 1105–25.
- Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–402.
- Andrássy, I. and L. Zombori (1976). *Evolution as a basis for the systematization of nematodes*. Pitman, London.
- Baillie, D. L. and A. M. Rose (2000). WABA success: a tool for sequence comparison between large genomes. *Genome Res.* 10, 1071–3.
- Baird, S., M. E. Sutherlin, and S. W. Emmons (1992). Reproductive isolation in Rhabditidae (Nematoda: Secernentea); mechanisms that isolate size species of three genera. *Evolution* 46, 585–94.
- Barnes, T. M., Y. Kohara, A. Coulson, and S. Hekimi (1995). Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. *Genetics* 141, 159–79.
- Benton, M. J. and F. J. Ayala (2003). Dating the tree of life. *Science* 300, 1698–700.
- Blair, J. E., K. Ikeo, T. Gojobori, and S. Blair Hedges (2002). The evolutionary position of nematodes. *BMC Evol. Biol.* 2, 7–14.
- Blanchette, M., T. Kunisawa, and D. Sankoff (1996). Parametric genome rearrangement. *Gene* 172, GC11–7.
- Blaxter, M. (1998). *Caenorhabditis elegans* is a nematode. *Science* 282, 2041–6.
- Blaxter, M. (2000). Genes and genomes of *Necator americanus* and related hookworms. *Int. J. Parasitol.* 30, 347–55.
- Blaxter, M. L., P. De Ley, J. R. Garey, L. X. Liu, P. Scheldeman, A. Vierstraete, J. R. Vanfleteren, L. Y. Mackey, M. Dorris, L. M. Frisse, et al. (1998). A molecular evolutionary framework for the phylum Nematoda. *Nature* 392, 71–5.
- Boeckmann, B., A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365–70.
- Bonen, L. and J. Vogel (2001). The ins and outs of group II introns. *Trends Genet.* 17, 322–31.
- Brenner, S. (1963). Proposal to the Medical Research Council. See <http://elegans.swmed.edu/Sydney.html>.



- Brooke, M. L. (1999). How old are animals? *Trends Ecol. Evol.* 14, 211–2.
- Burt, D. W., C. Bruley, I. C. Dunn, C. T. Jones, A. Ramage, A. S. Law, D. R. Morrice, I. R. Paton, J. Smith, D. Windsor, et al. (1999). The dynamics of chromosome evolution in birds and mammals. *Nature* 402, 411–3.
- Butler, M. H., S. M. Wall, K. R. Luehrsen, G. E. Fox, and R. M. Hecht (1981). Molecular relationships between closely related strains and species of nematodes. *J. Mol. Evol.* 18, 18–23.
- Cáceres, M., J. M. Ranz, A. Barbadilla, M. Long, and A. Ruiz (1999). Generation of a widespread *Drosophila* inversion by a transposable element. *Science* 285, 415–8.
- Carlton, J. M., S. V. Angiuoli, B. B. Suh, T. W. Kooij, M. Perlea, J. C. Silva, M. D. Ermolaeva, J. E. Allen, J. D. Selengut, H. L. Koo, et al. (2002). Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* 419, 512–9.
- Carmi, I., J. B. Kocpczynski, and B. J. Meyer (1998). The nuclear hormone receptor SEX-1 is an X-chromosome signal that determines nematode sex. *Nature* 396, 168–73.
- Cavalier-Smith, T. (1985). Selfish DNA and the origin of introns. *Nature* 315, 283–4.
- Charlesworth, B., J. A. Coyne, and N. H. Barton (1987). The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* 130, 113–46.
- Chitwood, B. G. and M. B. Chitwood (1974). *Introduction to Nematology*. University Park Press, Baltimore.
- Clamp, M., D. Andrews, D. Barker, P. Bevan, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, et al. (2003). Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* 31, 38–42.
- Cobb, N. A. (1915). Nematodes and their relationships. In *Year Book Dept. Agric. 1914*, pp. 457–90. Washington, DC: Dept. Agric.
- Coghlan, A. and K. H. Wolfe (2002). Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. *Genome Res.* 12, 857–67.
- Crick, F. (1979). Split genes and RNA splicing. *Science* 204, 264–71.
- Cutter, A. D. and B. A. Payseur (2003). Rates of deleterious mutation and the evolution of sex in *Caenorhabditis*. *J. Evol. Biol.* 16, 812–22.
- Dehal, P., P. Predki, A. S. Olsen, A. Kobayashi, P. Folta, S. Lucas, M. Land, A. Terry, C. L. Ecale Zhou, S. Rash, et al. (2001). Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* 293, 104–11.
- Dibb, N. J. and A. J. Newman (1989). Evidence that introns arose at proto-splice sites. *EMBO J.* 8, 2015–21.
- Eichler, E. E. and D. Sankoff (2003). Structural dynamics of eukaryotic chromosome evolution. *Science* 301, 793–7.
- Emmons, S. W., M. R. Klass, and D. Hirsh (1979). Analysis of the constancy of DNA sequences during development and evolution of the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* 76, 1333–7.
- Enright, A. J., S. Van Dongen, and C. A. Ouzounis (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–84.
- Evans, D., D. Zorio, M. MacMorris, C. Winter, K. Lea, and T. Blumenthal (1997). Operons and SL2 trans-splicing exist in nematodes outside the genus *Caenorhabditis*. *Proc. Natl. Acad. Sci. USA* 94, 9751–6.



- Fedorov, A., X. Cao, S. Saxonov, S. J. de Souza, S. W. Roy, and W. Gilbert (2001). Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. *Proc. Natl. Acad. Sci. USA* 98, 13177–82.
- Fedorov, A., S. Roy, X. Cao, and W. Gilbert (2003). Phylogenetically older introns strongly correlate with module boundaries in ancient proteins. *Genome Res.* 13, 1155–7.
- Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) versions 3.5c and 3.6a3. Department of Genetics, University of Washington, Seattle.
- Frugoli, J. A., M. A. McPeck, T. L. Thomas, and C. R. McClung (1998). Intron loss and gain during evolution of the catalase gene family in angiosperms. *Genetics* 149, 355–65.
- Gilbert, W. (1978). Why genes in pieces? *Nature* 271, 501.
- Giroux, M. J., M. Clancy, J. Baier, L. Ingham, D. McCarty, and L. C. Hannah (1994). *De novo* synthesis of an intron by the maize transposable element *Dissociation*. *Proc. Natl. Acad. Sci. USA* 91, 12150–4.
- Goff, S. A., D. Rieke, T.-H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296, 92–100.
- González, J., J. M. Ranz, and A. Ruiz (2002). Chromosomal elements evolve at different rates in the *Drosophila* genome. *Genetics* 161, 1137–54.
- Graustein, A., J. M. Gaspar, J. R. Walters, and M. F. Palopoli (2002). Levels of DNA polymorphism vary with mating system in the nematode genus *Caenorhabditis*. *Genetics* 161, 99–107.
- Gu, X. and J. Zhang (1997). A simple method for estimating the parameter of substitution rate variation among sites. *Mol. Biol. Evol.* 14, 1106–13.
- Guiliano, D. B., N. Hall, S. J. M. Jones, L. N. Clark, C. H. Corton, B. G. Barrell, and M. L. Blaxter (2002). Conservation of long-range synteny and microsynteny between the genomes of two distantly related nematodes. *Genome Biol.* 3, RESEARCH0057.1–0057.14.
- Haig, D. (1999). A brief history of human autosomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 354, 1447–70.
- Hankeln, T., H. Friedl, I. Ebersberger, J. Martin, and E. R. Schmidt (1997). A variable intron distribution in globin genes of *Chironomus*: evidence for recent intron gain. *Gene* 205, 151–60.
- Hannenhalli, S. (1996). Polynomial-time algorithm for computing translocation distance between genomes. *Discrete Applied Math.* 71, 137–51.
- Henikoff, S. and J. G. Henikoff (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89, 10915–9.
- Heschl, M. F. and D. L. Baillie (1990). Functional elements and domains inferred from sequence comparisons of a heat shock gene in two nematodes. *J. Mol. Evol.* 31, 3–9.
- Hill, A. A., C. P. Hunter, B. T. Tsung, G. Tucker-Kellogg, and E. L. Brown (2000). Genomic analysis of gene expression in *C. elegans*. *Science* 290, 809–12.
- Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, et al. (2002). The Ensembl genome database project. *Nucleic Acids Res.* 30, 38–41.



- Jovelin, R., B. C. Ajie, and P. C. Phillips (2003). Molecular evolution and quantitative variation for chemosensory behaviour in the nematode genus *Caenorhabditis*. *Mol. Ecol.* *12*, 1325–37.
- Kaestner, A. (1965). *Lehrbuch der Speziellen Zoologie. Band I: Wirbellose*. Verlag, Jena.
- Kececioğlu, J. and R. Ravi (1995). Of mice and men: Evolutionary distances between genomes under translocation. In *Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 604–13. SIAM, Philadelphia, PA.
- Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* *423*, 241–54.
- Kennedy, B. P., E. J. Aamodt, F. L. Allen, M. A. Chung, M. F. Heschl, and J. D. McGhee (1993). The gut esterase gene (*ges-1*) from the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *J. Mol. Biol.* *229*, 890–908.
- Kent, W. J. and A. M. Zahler (2000). Conservation, regulation, synteny, and introns in a large-scale *C. briggsae*-*C. elegans* genomic alignment. *Genome Res.* *10*, 1115–25.
- Kirouac, M. and P. W. Sternberg (2003). *cis*-Regulatory control of three cell fate-specific genes in vulval organogenesis of *Caenorhabditis elegans* and *C. briggsae*. *Dev. Biol.* *257*, 85–103.
- Korf, I., P. Flicek, D. Duan, and M. R. Brent (2001). Integrating genomic homology into gene structure prediction. *Bioinformatics* *17 Suppl. 1*, S140–8.
- Kuwabara, P. E. and S. Shah (1994). Cloning by synteny: identifying *C. briggsae* homologues of *C. elegans* genes. *Nucleic Acids Res.* *22*, 4414–8.
- Lande, R. (1979). Effective deme size during long-term evolution estimated from rates of chromosomal rearrangements. *Evolution* *33*, 234–51.
- Lee, J. M. and E. L. L. Sonnhammer (2003). Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* *13*, 875–82.
- Lee, K.-Z., A. Eizinger, R. Nandakumar, S. C. Schuster, and R. J. Sommer (2003). Limited microsynteny between the genomes of *Pristionchus pacificus* and *Caenorhabditis elegans*. *Nucleic Acids Res.* *31*, 2553–60.
- Lee, K.-Z. and R. J. Sommer (2003). Operon structure and trans splicing in the nematode *Pristionchus pacificus*. *Mol. Biol. Evol.* (in press).
- Lee, Y. H., X. Y. Huang, D. Hirsh, G. E. Fox, and R. M. Hecht (1992). Conservation of gene organization and trans-splicing in the glyceraldehyde-3-phosphate dehydrogenase-encoding genes of *Caenorhabditis briggsae*. *Gene* *121*, 227–35.
- Lercher, M. J., T. Blumenthal, and L. D. Hurst (2003). Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. *Genome Res.* *13*, 238–43.
- Lilley, C. J., P. Devlin, P. E. Urwin, and H. J. Atkinson (1999). Parasitic nematodes, proteinases and transgenic plants. *Parasitol. Today* *15*, 414–7.
- Logsdon, J. M. J. (1998). The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev.* *8*, 637–48.
- Logsdon, J. M. J., A. Stoltzfus, and W. F. Doolittle (1998). Molecular evolution: recent cases of spliceosomal intron gain? *Curr. Biol.* *8*, R560–3.



- Logsdon, J. M. J., M. G. Tyshenko, C. Dixon, J. D-Jafari, V. K. Walker, and J. D. Palmer (1995). Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc. Natl. Acad. Sci. USA* 92, 8507-11.
- Long, M. (2001). Evolution of novel genes. *Curr. Opin. Genet. Dev.* 11, 673-80.
- Long, M., S. J. de Souza, C. Rosenberg, and W. Gilbert (1998). Relationship between "proto-splice" sites and intron phases: evidence from dicodon analysis. *Proc. Natl. Acad. Sci. USA* 95, 219-23.
- Luong, T. V. (2003). De-worming school children and hygiene intervention. *Int. J. Environ. Health Res.* 13 Suppl 1, S153-9.
- Lynch, M. (2002). Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. USA* 99, 6118-23.
- Mironov, A. A., M. A. Roytberg, P. A. Pevzner, and M. S. Gelfand (1998). Performance-guarantee gene predictions via spliced alignment. *Genomics* 51, 332-9.
- Mushegian, A. R., J. R. Garey, J. Martin, and L. X. Liu (1998). Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. *Genome Res.* 8, 590-8.
- Nadeau, J. H. and B. A. Taylor (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. USA* 81, 814-8.
- Nesterova, T. B., S. M. Duthie, N. A. Mazurok, A. A. Isaenko, N. V. Rubtsova, S. M. Zakian, and N. Brockdorff (1998). Comparative mapping of X chromosomes in vole species of the genus *Microtus*. *Chromosome Res.* 6, 41-8.
- Nigon, V. and E. C. Dougherty (1949). Reproductive patterns and attempts at reciprocal crossing of *Rhabditis elegans* Maupas, 1900, and *Rhabditis briggsae* Dougherty and Nigon, 1949 (Nematoda: Rhabditidae). *J. Exp. Zool.* 112, 485-503.
- Notredame, C., D. G. Higgins, and J. Heringa (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302, 205-17.
- Ohno, S. (1967). Sex chromosomes and sex-linked genes. In Labhart, A. et al. (Ed.), *Monographs on endocrinology*, Volume 1, pp. 123-35. Springer-Verlag, Heidelberg.
- O'Neill, R. J., F. E. Brennan, M. L. Delbridge, R. H. Crozier, and J. A. Graves (1998). *De novo* insertion of an intron into the mammalian sex determining gene, *SRY*. *Proc. Natl. Acad. Sci. USA* 95, 1653-7.
- Oosumi, T., B. Garlick, and W. R. Belknap (1995). Identification and characterization of putative transposable DNA elements in solanaceous plants and *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* 92, 8886-90.
- Oosumi, T., B. Garlick, and W. R. Belknap (1996). Identification of putative nonautonomous transposable elements associated with several transposon families in *Caenorhabditis elegans*. *J. Mol. Evol.* 43, 11-8.
- Pearson, W. R. and D. J. Lipman (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85, 2444-8.
- Pevzner, P. and G. Tesler (2003a). Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.* 13, 37-45.
- Pevzner, P. and G. Tesler (2003b). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. USA* 100, 7672-7.



- Poinar, Jr., G. O. (1983). *The natural history of nematodes*. Prentice-Hall, Englewood Cliffs, NJ.
- Prasad, S. S. and D. L. Baillie (1989). Evolutionarily conserved coding sequences in the *dpy-20-unc-22* region of *Caenorhabditis elegans*. *Genomics* 5, 185–98.
- Purugganan, M. D. (2002). The splicing of transposable elements: evolution of a nuclear defense against genomic invasions? In M. Syvanen and C. I. Kado (Eds.), *Horizontal Gene Transfer*, pp. 187–95. Chapman-Hall, London.
- Ranz, J. M., F. Casals, and A. Ruiz (2001). How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res.* 11, 230–9.
- Ranz, J. M., J. González, F. Casals, and A. Ruiz (2003). Low occurrence of gene transposition events during the evolution of the genus *Drosophila*. *Evolution* 57, 1325–35.
- Reboul, J., P. Vaglio, J.-F. Rual, P. Lamesch, M. Martinez, C. M. Armstrong, S. Li, L. Jacotot, N. Bertin, R. Janky, et al. (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.
- Reese, M. G., G. Hartzell, N. L. Harris, U. Ohler, J. F. Abril, and S. E. Lewis (2000). Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* 10, 483–501.
- Riddle, D. L., T. Blumenthal, B. J. Meyer, and J. R. Priess (Eds.) (1997). *C. elegans II*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Robertson, H. M. (2001). Updating the *str* and *srj* (*stl*) families of chemoreceptors in *Caenorhabditis* nematodes reveals frequent gene movement within and between chromosomes. *Chem. Senses* 26, 151–9.
- Rogers, J. H. (1989). How were introns inserted into nuclear genes? *Trends Genet.* 5, 213–6.
- Rogic, S., B. F. F. Ouellette, and A. K. Mackworth (2002). Improving gene recognition accuracy by combining predictions from two gene-finding programs. *Bioinformatics* 18, 1034–45.
- Rogozin, I. B., J. Lyons-Weiler, and E. V. Koonin (2000). Intron sliding in conserved gene families. *Trends Genet.* 16, 430–2.
- Rogozin, I. B., Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, and E. V. Koonin (2003). Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* 13, 1512–7.
- Romero, D., J. Martinez-Salazar, E. Ortiz, C. Rodriguez, and E. Valencia-Morales (1999). Repeated sequences in bacterial chromosomes and plasmids: a glimpse from sequenced genomes. *Res. Microbiol.* 150, 735–43.
- Roy, P. J., J. M. Stuart, J. Lund, and S. K. Kim (2002). Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature* 418, 975–9.
- Roy, S. W. (2003). Recent evidence for the exon theory of genes. *Genetica* 118, 251–66.
- Roy, S. W., A. Fedorov, and W. Gilbert (2003). Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl. Acad. Sci. USA* 100, 7158–62.
- Salamov, A. A. and V. V. Solovyev (2000). Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* 10, 516–22.



- Sankoff, D. (1999). Comparative mapping and genome rearrangement. In J. C. M. Dekkers, S. J. Lamont, and M. F. Rothschild (Eds.), *From Jay Lush to genomics: Visions for animal breeding and genetics*, pp. 124–34. Iowa State University, Ames, IA.
- Sharakhov, I. V., A. C. Serazin, O. G. Grushko, A. Dana, N. Lobo, M. E. Hillenmeyer, R. Westerman, J. Romero-Severson, C. Costantini, N. Sagnon, et al. (2002). Inversions and gene order shuffling in *Anopheles gambiae* and *A. funestus*. *Science* 298, 182–5.
- Sharp, P. A. (1985). On the origin of RNA splicing and introns. *Cell* 42, 397–400.
- Sivasundar, A. and J. Hey (2003). Population genetics of *Caenorhabditis elegans*. The paradox of low polymorphism in a widespread species. *Genetics* 163, 147–57.
- Stein, L., P. Sternberg, R. Durbin, J. Thierry-Mieg, and J. Spieth (2001). WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.* 29, 82–6.
- Stein, L. D., Z. Bao, D. Blasiar, T. Blumenthal, M. Brent, N. Chen, A. Chinwalla, L. Clarke, C. Clee, A. Coghlan, et al. (2003). The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biology* (in press).
- Stothard, P. and D. Pilgrim (2003). Sex-determination gene and pathway evolution in nematodes. *Bioessays* 25, 221–31.
- Surzycki, S. A. and W. R. Belknap (2000). Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. *Proc. Natl. Acad. Sci. USA* 97, 245–9.
- Takacs, A. M., J. A. Denker, K. G. Perrine, P. A. Maroney, and T. W. Nilsen (1988). A 22-nucleotide spliced leader sequence in the human parasitic nematode *Brugia malayi* is identical to the trans-spliced leader exon in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* 85, 7932–6.
- Takezaki, N., A. Rzhetsky, and M. Nei (1995). Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.* 12, 823–33.
- Tarrío, R., F. Rodríguez-Trelles, and F. J. Ayala (1998). New *Drosophila* introns originate by duplication. *Proc. Natl. Acad. Sci. USA* 95, 1658–62.
- Tarrío, R., F. Rodríguez-Trelles, and F. J. Ayala (2003). A new *Drosophila* spliceosomal intron position is common in plants. *Proc. Natl. Acad. Sci. USA* 100, 6580–3.
- Thacker, C., M. A. Marra, A. Jones, D. L. Baillie, and A. M. Rose (1999). Functional genomics in *Caenorhabditis elegans*: An approach involving comparisons of sequences from related nematodes. *Genome Res.* 9, 348–59.
- The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282, 2012–8.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson (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, 4673–80.
- Timakov, B., X. Liu, I. Turgut, and P. Zhang (2002). Timing and targeting of *P*-element local transposition in the male germline cells of *Drosophila melanogaster*. *Genetics* 160, 1011–22.
- Vanfleteren, J. R., Y. Van de Peer, M. L. Blaxter, S. A. Tweedie, C. Trotman, L. Lu, M. L. Van Hauwaert, and L. Moens (1994). Molecular genealogy of some nematode taxa as based on cytochrome c and globin amino acid sequences. *Mol. Phylogenet. Evol.* 3, 92–101.



- Venkatesh, B., Y. Ning, and S. Brenner (1999). Late changes in spliceosomal introns define clades in vertebrate evolution. *Proc. Natl. Acad. Sci. USA* 96, 10267–71.
- Wang, D. Y., S. Kumar, and S. B. Hedges (1999). Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B Biol. Sci.* 266, 163–71.
- Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–62.
- Wootton, J. C. and S. Federhen (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* 266, 554–71.
- Zdobnov, E. M., C. von Mering, I. Letunic, D. Torrents, M. Suyama, R. R. Copley, G. K. Christophides, D. Thomasova, R. A. Holt, G. M. Subramanian, et al. (2002). Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* 298, 149–59.
- Zhang, J. and T. Peterson (1999). Genome rearrangements by nonlinear transposons in maize. *Genetics* 153, 1403–10.