Conservation, Regulation, Synteny, and Introns in a Large-scale *C. briggsae–C. elegans* Genomic Alignment

W. James Kent[1] and Alan M. Zahler

Department of Biology and Center for Molecular Biology of RNA, University of California at Santa Cruz, Santa Cruz, California 95064 USA

A new algorithm, WABA, was developed for doing large-scale alignments between genomic DNA of different species. WABA was used to align 8 million bases of *Caenorhabditis briggsae* genomic DNA against the entire 97-million-base *Caenorhabditis elegans* genome. The alignment, including *C. briggsae* homologs of 154 genetically characterized *C. elegans* genes and many times this number of largely uncharacterized ORFs, can be browsed and searched on the Web (http://www.cse.ucsc.edu/~kent/intronerator). The alignment confirms that patterns of conservation can be useful in identifying regulatory regions and rarely expressed coding regions. Conserved regulatory elements can be identified inside coding exons by examining the level of divergence at the wobble position of codons. The alignment reveals a bimodal size distribution of syntenic regions. Over 250 introns are present in one species but not the other. The 3′ and 5′ intron splice sites have more similarity to each other in introns unique to one species than in *C. elegans* introns as a whole, suggesting a possible mechanism for intron removal.

Biologists often compare DNA sequences to determine the relationships between species and to learn about the function of genes. As a result of the efforts of various sequencing projects, it is now becoming possible to do sequence comparisons between complete genomes. These comparisons can answer questions about how genomes change over time in more detail than is possible by chromosome painting (O'Brien et al. 1999). Comparing coding regions between species can locate functionally important parts of proteins, which are more highly conserved than other parts (Makalowski et al. 1996; Makalowski and Boguski 1998). Comparing noncoding regions of homologous genes between species separated by an appropriate evolutionary distance can locate promoters and other conserved regulatory elements (Hardison et al. 1997; Endrizzi et al. 1999; Jareborg et al. 1999). Comparisons between mouse and human, between maize and rice (Wilson et al. 1999), and between *Fugu rubripes*, *Tetraodon fluviatilis*, and *Danio rerio* (Boeddrich et al. 1999) have all revealed important functional regions of the genome.

This paper presents a genomic DNA comparison between the nematodes *Caenorhabditis briggsae* and *Caenorhabditis elegans*. *C. elegans* and *C. briggsae* are closely related nematodes in the same genus (Baldwin et al. 1997; Blaxter 1998; Blaxter et al. 1998; Voronov et al. 1998). They are estimated to have diverged 25–50 million years ago, although without a fossil record these estimates are highly dependent on assumptions about mutation rates that vary considerably between organisms and between genes (Ayala et al. 1996). In practical terms, *C. elegans* and *C. briggsae* are separated by a nearly ideal distance for comparative genomics. In regions experiencing selective pressure, close to 80% base identity is preserved between species. In other regions base identity is close to 30% (Shabalina and Kondrashov 1999). The large-scale comparative genomic study we present here aligns 8 million bases of *C. briggsae* sequence from 229 cosmids with 97 million bases of *C. elegans* sequence covering essentially the entire genome (*C. elegans* Sequencing Consortium 1998). The scale of this comparison presented unique challenges and resulted in the development of new algorithms that can cope with long insertions. Our algorithms are also able to recognize homologous regions at the DNA level despite the rapid divergence in the wobble position of most codons.

[1]Corresponding author.
E-MAIL kent@biology.ucsc.edu; FAX (831) 459–3737.

RESULTS AND DISCUSSION

An Algorithm for Interspecies Genomic Sequence Alignments

The alignment of 8 million bases of *C. briggsae* against 97 million bases of *C. elegans* involves, at least conceptually, almost 800 trillion nucleotide comparisons. This requires a relatively fast algorithm. Genomic alignment also requires handling both small and large insertions and deletions and a high degree of divergence in the third, wobble, position of codons. We developed a three-pass algorithm—the Wobble Aware Bulk Aligner (WABA)—that meets these requirements

10:1115–1125 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

and is described in some detail in Methods. Homologous regions were identified with the first pass of WABA, which took 20 hr of run time on a Pentium III 450 mHz. The second pass, which did a detailed alignment of overlapping 2000 × 5000 base regions, took 11 days to run on the same machine. The third pass joined the overlapping alignments in 15 min on a comparable machine. Although the first and especially the second passes are slow on a single workstation, it is easy to distribute them on many machines. Fortunately, it is in the first pass that the run time varies with the product of the genome sizes that are being compared, where the run times of the other passes are close to linear with respect to the shorter genome size. It should thus be possible in the future to apply WABA to aligning large vertebrate genomes.

To help assess the utility of WABA, we performed the same alignment with BLAST (Altschul et al. 1997). We ran the BLASTN program with two different settings: the default and the slower, but more sensitive, settings. The results are summarized in Table 1. WABA aligned over twice as many bases as BLAST. The average length of a single WABA alignment was >10 times as long as the average BLAST alignment. WABA alignments with inserts of >5 bases or with sections of poor conservation >20 bases long tended to appear as several separate BLAST alignments. Blast found 80% of the regions WABA classified as coding or highly conserved (see below). Table 2 shows an alignment that only WABA found. Running BLAST with a window size of 8 and an extension threshold of 8 rather than the default 11 and 11 resulted in BLAST aligning 6% more bases, but this also increased the run time by a factor of 41. WABA was also slower than BLAST at default settings but only by a factor of 24.

Web-based Browser and Database

The results of the genomic sequence alignment of *C. briggsae* with *C. elegans* are available through a Web-based display called the Intronerator. The alignments can be viewed both as a graphical display indicating the degree of sequence conservation and as a base-by-base alignment in the context of gene predictions and mRNA alignments (Kent and Zahler 2000) at http://www.cse.ucsc.edu/~kent/intronerator. A sample Intronerator high-level view is shown in Figure 1. This high-level view displays the region of the *C. elegans* genome containing the *unc-47* gene. Regions containing homology to *C. briggsae* are represented by a bar immediately under the gene predictions. Areas of stronger homology are represented in a darker shading of the bar than regions of lower homology. Clicking on the bar brings up a base-by-base view of the alignment. The browser is entered via the Tracks Display link on the Intronerator home page. It is possible to search as well as browse the alignments by following the Extract Sequences link. *C. elegans* genes can be searched for ORFs, for sequenced genes, for the presence of aligning mRNA, and for the presence of aligning *C. briggsae* genomic sequence. The various search criteria can be combined in a flexible manner. The names of genetically characterized and sequenced *C. elegans* genes with *C. briggsae* sequence homology are shown in Table 3. The program can also restrict the returned sequences to introns only, exons only, or windows relative to the translational start or finish. This last feature can be used to extract likely promoter regions. A Web interface to the alignment algorithm itself is available at http://www.cse.ucsc.edu/~kent/xenoAli/. From this page the user can align sequences against the *C. elegans* genome or align two sequences to each other. The page also links to a list of all homologous *C. briggsae* and *C. elegans* regions sorted by either alignment score, location in *C. briggsae*, or location in *C. elegans*.

Alignment Statistics

Overall, 59% of the *C. briggsae* sequence is homologous to *C. elegans*. The alignment algorithm developed here

Table 1. Comparison of BLASTN and WABA Alignments

	BLASTN default	BLASTN −W8 −f8	WABA
Total base matches	1849727	1969827	4697347
Total base mismatches	314023	340606	2538033
Total bases inserted	5859	5659	2275269
Total number of inserts	2836	2962	27469
Number of alignments	14653	16247	3839
Average size of alignments	147.9	142.4	2059.3
C. briggsae cosmids total	229	229	229
Cosmids with any alignment	224	224	227
Seconds to align one cosmid	60	2461	1420

Only alignments containing 50 or more base matches are included here. BLASTN was run at the default settings and with the window size and the minimum extension score set down to eight for increased sensitivity. WABA is more tolerant of long inserts and nucleotide mismatches than BLASTN at either setting and consequently is able to make longer alignments.

Table 2. An Alignment WABA Finds but BLAST Misses Between the *C. briggsae* Cosmid G47H01 and the *C. elegans* Cosmid T23C6

```
caaatgacaaggatagtttgaattcatccaaagataagaccaactacgcc
||  |  ||||||| | ||||||||| || || || || ||||| |||||
cagaccacaaggacaatttgaattcttcaaaggacaaaaccaattacgcg

tcatcgaagagccgcttaaaccagaggatcatgagagagaagctgaagag
|| || || ||| | ||||| ||||| ||  |||| ||  || |||| ||
tcgtcaaaaagcagattaaatcagagaattgtgagggaacagttgaaaag

tgactgttgggattgttggaaagatattaatactatgatggttcaaggag
||||||  |||  |||||||||    ||||| || ||||||| |||||||
tgactgctgggtatgttggaagagcattaacacaatgatggtccaaggag
```

This is in an area predicted to be coding in *C. elegans.* Note that many of the mismatches occur either 3 or 6 nucleotides away from the previous mismatch indicating variation in the wobble codon.

has at its core a pair–hidden Markov model (HMM) (Durbin et al. 1998) that, along with the alignment itself, classifies regions into weakly conserved regions, strongly conserved regions, and conserved coding regions (see Methods). Within weakly conserved regions, an average of 49% of nucleotides are an identical match. Within strongly conserved regions, 89% of nucleotides match. In the first two nucleotide positions inside codons in conserved coding regions, 92% of nucleotides match. In the third, wobble, position 53% of the nucleotides match. Because, as discussed below, the discrimination between coding and highly conserved regions is not robust, for purposes of further analysis we lump the coding regions in with the highly conserved regions. Of the 8,074,467 bases of *C. briggsae* sequence, 30.5% are in strongly conserved regions, 28.5% are in weakly conserved regions, and 41.0% are in nonaligning regions. Fifty-five percent of the highly conserved regions are coding, 20% are in introns, and 25% are in intergenic regions according to the gene predictions of the *C. elegans* Sequencing Consortium (1998). It is likely that many of the conserved noncoding regions are regulatory in nature.

Compared with mammalian genomes in which ~50% of the bases are part of repeating elements (Smit 1999), there are relatively few repeating elements in *C. elegans* and *C. briggsae*. Nineteen percent of the strongly conserved regions and 8% of the weakly conserved *C. briggsae* regions align to more than one place in the *C. elegans* genome. Table 4 presents a breakdown of the highly conserved multiply aligning regions. The highly conserved intergenic and intronic sequences are less likely to align to multiple places than the coding bases, indicating that the repeating regions are dominated by duplicated genes. The region that aligned in the most places, 32 in all, was a coding region of the major sperm protein. The transposable element Tc*1* aligned in 25 places. The majority of repeating regions, however, only align to two or three places.

In the current implementation, WABA discriminates weakly between strongly conserved and coding regions. In the regions of alignment, the gene predictions of the *C. elegans* Sequencing Consortium (1998) classify 1,364,947 bases as coding, WABA classifies 1,845,169 bases as coding, and 1,192,602 bases are classified as coding by both WABA and the gene predictions. If the gene predictions are correct, then WABA has a false-negative rate of 13% and a false-positive rate of 55%, which is not as good as the current state of the art in gene prediction software. The goal of this project was merely to create a sensitive alignment tool, not a gene-finding program. However, in the future it would be interesting to extend the pair–HMM in such a way that it can model codon bias in general and stop codons in particular as well as account for wobble position variations. It is possible such extensions would yield a program that would produce both sensitive genomic alignments and reliable predictions of coding regions.

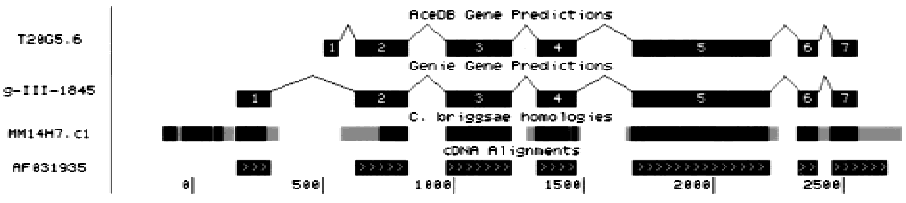

Figure 1 An Intronerator tracks display of the region around *unc-47*. The top track shows the gene prediction of the *C. elegans* Sequencing Consortium (1998) that is used in AceDB. The second track shows a prediction by the Genie gene finding program (Kulp et al. 1996, 1997; Reese et al. 1997). The third track gives a high level view of the *C. briggsae* homolog. Regions of high homology (classified by the seven state aligner as highly conserved or coding) appear darker. Regions of lower homology appear in a lighter shade. Nonaligning regions are white. The fourth track shows a cDNA alignment. Note that in this case the Genie prediction agrees with the mRNA, whereas the AceDB prediction misplaces the first exon. Note too the highly conserved region upstream of the first exon containing known promoter elements for this gene (Eastman et al. 1999).

Table 3. A List of Genetically Characterized and Sequence *C. elegans* Genes for Which There Is Homology in the Washington Univeristy *C. briggsae* Sequence Data

ace-1 ace-3 ace-4 act-4 add-2 avr-15 bli-4 cbp-1 ceh-6 cha-1 cmk-1 col-2 col-8 cpr-5 csr-1 daf-18 deg-3 del-1 des-2 dhc-1 dom-3 drp-1 eat-6 eft-2 egl-9 exp-2 fat-3 fem-1 fem-2 flp-13 gas-1 gcy-13 gcy-33 gcy-4 gcy-6 glc-3 glp-1 goa-1 gpb-1 gpd-1 gpd-2 gpd-3 gpd-4 her-1 his-10 his-11 his-12 his-29 his-31 his-32 his-33 his-34 his-35 his-38 his-5 his-6 his-7 his-8 his-9 hsb-1 hum-3 ife-4 ima-3 itr-1 jnk-1 kin-13 kin-15 kin-16 klc-1 klc-2 kup-1 lec-2 lec-3 let-2 let-70 let-721 let-805 let-858 let-99 lev-1 lgx-1 lin-12 lin-2 lin-25 lin-3 lin-46 lrp-1 lrx-1 mab-3 mai-1 mec-10 mec-12 mec-9 mel-32 mix-1 mom-5 msp-113 msp-142 msp-19 msp-31 msp-32 msp-33 msp-38 msp-51 msp-53 mua-3 mxl-1 myo-1 myo-2 myo-3 ncc-1 nhr-14 nhr-21 nhr-8 nid-1 par-3 pgp-2 pgp-3 pgp-4 rab-3 rrp-1 sel-12 sex-1 skn-1 sma-3 sma-6 spe-9 sqt-1 sqv-3 syd-2 tax-2 tba-1 tba-2 tnc-1 twk-11 twk-17 unc-115 unc-17 unc-22 unc-24 unc-30 unc-33 unc-43 unc-45 unc-47 unc-5 uvt-1 vab-8 vha-1 vha-2 vit-1 vit-2 vit-5 zen-4

Conservation of Promoters

Figure 1 shows an Intronerator display of the *unc-47* gene. This is a fairly typical example of the patterns of conservation in protein-coding genes between *C. elegans* and *C. briggsae*. Most of the coding regions are heavily conserved as is the promoter region. Beyond the splice consensus sequences, most introns are lightly conserved if at all. In this case the gene has been cloned and sequenced (McIntire et al. 1997) and the promoter region studied (Eastman et al. 1999).The conserved region upstream of the first exon covers almost exactly the minimum promoter needed to drive reporter constructs in the same expression pattern as *unc-47* (Eastman et al. 1999). Identification of conserved regions in promoters will help biologists identify the minimal promoter sequence necessary to drive the proper expression of a gene in functional studies.

Table 4. Distribution of Highly Conserved Regions Comparing the Number of Places a Region Aligns vs. Whether the Region is Coding, Intronic, or Intergenic

Places aligned	Coding bases	Intronic bases	Intergenic bases	Total bases
1	985406	434809	585204	2005419
2–3	214411	31393	27843	273647
4–7	84001	7957	3286	95244
8–15	54318	8482	609	63409
16–31	25687	2752	0	28439
32	124	0	0	124
Totals	1363947	484393	616942	2465282

The table lists the number of *C. briggsae* bases in each category. Coding/Intron/Intergenic classification is based on the *C. elegans* Sequencing Consortium (1998) gene predictions for the corresponding *C. elegans* regions. Eighty-one percent of the highly conserved bases align to a unique part of the *C. elegans* genome.

Cross-species alignments will be important for dissection of regulatory networks of transcription factors and their binding sites. Two other very powerful techniques for dissecting promoter logic are clustering expression patterns from DNA microarrays (Eisen et al. 1998) and searching for motifs with tools like MEME (Bailey and Elkan 1995) at http://www.sdsc.edu/MEME/meme/website and Improbizer at http://www.cse.ucsc.edu/~kent/improbizer/index.html. Together, these techniques can find *cis*-acting regions in the promoters of coexpressed genes. Areas in the promoter region conserved in cross-species alignments can serve as an independent confirmation that these regions are biologically relevant or, alternatively, can serve to reduce the size of the region that the motif searching algorithms must consider. As more microarray data is publicly released, it will become possible to apply these powerful bioinformatic approaches to characterizing transcription factor binding sites in *C. elegans*.

Conservation of Splicing Regulatory Elements

In genes that are alternatively spliced, such as *let-2* (Sibley et al. 1993), it is not unusual to find highly conserved areas in introns (Fig. 2). Presumably, these conserved areas may be involved in the regulation of alternative splicing. There are several blocks of conserved intron sequence in the alternatively spliced region of *let-2*. These are seen in the intron just upstream of exon 9, in the intron sequence between exons 9 and 10, and in several blocks between exons 10 and 11. A poly-GT motif conserved in the intron between exons 10 and 11 is rare in *C. elegans* introns as a whole. Searching the Intronerator's Alternative Splicing Catalog (Kent and Zahler 2000) shows that this motif does occur frequently in introns near alternative splice sites (data not shown). This poly-GT motif is also found near the region of alternative splicing of *bli-4* and is conserved between *C. briggsae* and *C. elegans* in that gene as well (Thacker et al. 1999).

Splicing regulatory elements are often found in exons as well. Because exons tend to be highly conserved, these regulatory elements do not stand out well in the high-level view. However, the base-by-base alignment view can be used to look for sequence conservation in the wobble positions inside coding regions, which only match 53% of the time in the alignment as a whole. Table 5 shows a base-by-base alignment of the *let-2* mutually exclusive exons 9 and 10. Sixty-eight bases at the end of exon 10 are perfectly conserved. This includes 22 wobble-position nucleotides. The last 19 bases of this conserved region are all purines. Some members of the SR protein splicing factors have been shown to bind directly to purine-rich elements in exons to promote regulated splicing events (Nagel et al. 1998)

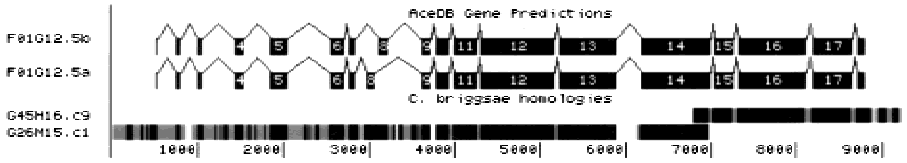

Figure 2 Tracks display of *let-2*, a gene that is alternatively spliced in a developmentally regulated manner. The *C. briggsae* sequence is split between two overlapping cosmids. Note the relatively high level of conservation in the introns compared with *unc-47*. (The exons numbered 8 on the top and bottom splicing diagrams are numbered 9 and 10 in the text and in Table 5. This difference is due to Tracks display not including exons in the UTR regions and numbering the exons in each isoform separately.)

Intron/Exon Structure of Rarely Expressed Genes

Even though there are > 100,000 *C. elegans* EST sequences in GenBank, only 9313 of 19,722 (47%) of the predicted *C. elegans* ORFs align with any ESTs. However, rarely expressed genes as well as commonly expressed genes can be identified by their patterns of conservation, as illustrated in Figure 3. Here, a single homologous region of *C. briggsae* covers two predicted *C. elegans* genes. One prediction is supported by seven EST sequences derived from four different cDNA clones. The other gene prediction has no supporting EST sequences at all. However, both genes are strongly supported by the pattern of conservation of predicted exons in the homologous *C. briggsae* region.

Introns Unique to One Species

Long insertions and deletions (indels) are common in the alignments. For indels from 3 to 30 nucleotides long, indels of an integral multiple of three were more common than other indels (data not shown). As indicated in Table 6, 95% of the longer indels occurred in regions of weak homology. Of the indels in regions of strong homology, 72% were in coding regions. Seventy-three percent of the indels in coding regions of strong homology, 263 in all, aligned with GT . . . AG ends on the region unique to one species. These indels appear to correspond to introns present in one species but not the other (Table 7). 263 unique introns is a conservative number. There are many additional cases where it appears that a codon or two were inserted or deleted along with the intron. In addition, there are cases where the insert resembles an intron that has since undergone a few point mutations in the flanking region. There also may be introns inserted or deleted in weakly conserved regions and noncoding regions as well. The figure of 263 unique introns does not include any of these cases.

From an alignment between just two sequences it is not possible to tell whether the introns are being gained, lost, or a mixture of both. There is only a small amount of genomic sequence data available for other members of the genus *Caenorhabditis* in GenBank and only a single instance (Hansen and Pilgrim 1998) where this data covers a coding indel in *C. elegans* and *C. briggsae*. In this case *C. briggsae* is missing a stretch of DNA that is present in *C. elegans* and *Caenorhabditis remanei*. *C. remanei* is considered to be a closer relative of *C. briggsae* than of *C. elegans* (Blaxter et al. 1998). Thus, this indel appears to represent an intron that was

Table 5. Alternatively Sliced Region of *let-2*

```
            aatttttgtattttcaaaattctagggagatctaggtttggtcggaccac
            |||| |||||||||| |||||||||||||||||||||  ||||||||||
            aattattgtattttctaaattctagGGAGATCTAGGTTCTGTCGGACCAC

            caggtccaccaggaccacgtgagttcactggatccggctcgattgtcgga
            |||||||||| |||||||||||||||| |||||||||| ||||||||| exon 9
            CAGGTCCACCGGGACCACGTGAGTTCACCGGATCCGGCTCAATTGTCGGA

            ccacgcggaaaccctggagaaaagggagacaaggtacttggtttattaat
            || |||||||||||||||||||||||||||||||||| | |||||||||
            CCTCGCGGAAACCCTGGAGAAAAGGGAGACAAGgtact-g-tttattaat

            tttttccatttt-agggagacattggtggcatgggaccggccggtccacc
            | ||||  |||| ||||||||||||||  ||||||||||||||||| ||
            tctttctttttttagGGAGACATTGGTGCGATGGGACCGGCCGGTCCGCC

            aggcccgatcgcctccaccatgtccaagggaaccattatcggtcctaagg
            |||||| |||||||||||||||||||||||||||||||||||||||||| exon10
            AGGCCCAATCGCCTCCACCATGTCCAAGGGAACCATTATCGGTCCTAAGG

            gagacctaggagagaagggagagaaggcaagaatatgtagctagtactaa
            |||||||||||||||||||||||||||||||||  |||      ||||
            GAGACCTAGGAGAGAAGGGAGAGAAGgcaagaatacttagagcaggctaa
```

C. briggsae is on top; *C. elegans* is below. Coding regions in *C. elegans* appear in upper case. Note the striking degree of conservation of the end of exon 10 and the unusual intron beginning with GC rather than the canonical GT that follows.

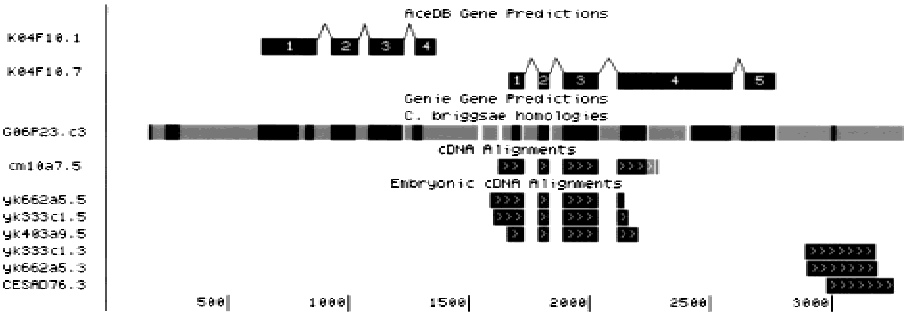

Figure 3 *C. elegans* predicted genes K04F10.1 and K04F10.7. The K04F10.7 gene prediction is supported by a number of ESTs. K04F10.1 appears by lack of EST coverage to be rarely expressed. The pattern of conservation of K04F10.1 nonetheless provides strong support for this gene prediction.

lost along with a single additional codon sometime since *C. briggsae* and *C. remanei* diverged. In a previous study comparing sequences of two large families of chemoreceptor genes in *C. elegans* and *C. briggsae* (Robertson 1998), Robertson found a pervasive pattern of intron loss over recent evolutionary time. Therefore, in all, it seems most likely that the majority of the introns unique to one species represent intron loss in the other species. Robertson noted a bias toward introns being lost in *C. briggsae* rather than *C. elegans*. Assuming the unique introns we observe in this larger study represent intron loss, 65% of the introns were lost from *C. briggsae* (Table 6).

We tested the model that flanking the unique introns may be some sequences associated with intron loss. The consensus sequence near the splice site of unique introns is different from the consensus sequence near the splice site of *C. elegans* introns as a whole (Table 8). In particular, the consensus of the AG before the 5′ splice site is much stronger in unique introns, particularly in *C. briggsae*. This and other changes to the consensus have the effect of strengthening the homology between the 3′ and 5′ splice sites. However, the homology is only strengthened significantly in the two bases before, and the one base after, the splice site (Table 9). This homology is too short for homologous recombination to be the mechanism for intron loss. However, the Ku mechanism for repairing double-stranded breaks in DNA (Jeggo 1998) joins broken ends at regions of microhomology (≈3 bp). The Ku mechanism deletes bases between the broken ends and the microhomology region as well. Thus, it seems possible that introns may be lost during repair of double-stranded breaks in DNA.

Table 6. Summary of Insertions and Introns Unique to One Species

11189	total inserts of length >32
6845	*C. elegans* inserts
4344	*C. briggsae* inserts
10686	inserts in regions of weak homology
503	inserts in regions of strong homology
4644	inserts in intergenic regions
4517	inserts in intron regions
1508	inserts in coding regions
518	inserts with one end in an intron and the other in an exon
2	inserts with one end in a gene and one end intergenic
301	inserts with GT. . .AG ends in regions of strong homology
362	inserts in coding regions with strong homology
170	introns unique to *C. elegans*
93	introns unique to *C. briggsae*
263	total introns unique to either species

An intron unique to one species is defined here as an insert of length >32 in a highly conserved coding region that aligns optimally with GT . . . AG ends.

Synteny

The large scale of this alignment also permitted us to observe the level of short-range synteny between the two genomes. On average, a *C. briggsae* sequence would only align for 8553 bases to one part of the *C. elegans* genome before the best alignment shifted to another part of the *C. elegans* genome. Because the average size of a *C. briggsae* clone was only 35,722 bases, approximately one-fourth of the alignments ended because the *C. briggsae* sequence ended. Thus, the average size of an aligning segment is actually larger. Figure 4 depicts the number of fragments clones are broken into by the alignment. Curiously, this quantity exhibits a bimodal distribution, with peaks at one fragment per clone and three fragments per clone. The first peak indicates that ~40% of the genome is resistant to rearrangment, going longer between rearrangement events than the 35,722-bp size of the *C. elegans* clones. The other 60% of the genome appears to be susceptible to rearrangements, with rearrangement events occurring approximately every 4000 bases. Because the average length between genes in *C. elegans* is ~5000 nucleotides, this is consistent with arrangement occurring freely in regions between genes in the 60% of the genome that is susceptible to rearrangement.

The 40% of the genome resistant to rearrangement may reflect regions that are transcriptionally coregulated. The largest single alignment is part of the homeobox complex. Genes that are part of operons (Spi-

Table 7. Examples of Introns Unique to One Species

A	cccaatccaagagaggtaactattttt	...	taattctttttcaggtatcctaactctc	*briggsae*
	cccagtccaagagag------------	...	--------------gtatccttaccctc	*elegans*
B	agaatgtctgtagta------------	...	--------------gaaatgaacaagcg	*briggsae*
	agaaattcgttggtggtaagctttaca	...	ataaataaattcaggaaatgaacaaacg	*elegans*
C	ctacatttttcattggttcgttttcca	...	gagtcgctttcagttaaa-aagtat-tt	*briggsae*
	ccactttctttattg------------	...	--------------taaagaagtatctt	*elegans*
D	agttttcgagatcacacaggtaaccgt	...	atgatcttttgattccagcacatctccc	*briggsae*
	agttttcgagatcacCcag--------	...	------------------cacatttctc	*elegans*

(A,B) Two of the 263 inserts in highly conserved coding regions that align optimally with the GT . . . AG ends characteristic of an intron. (C,D) Other inserts in coding regions that are potentially introns that have been inserted or deleted and have undergone additional mutation as well. (C) An insert that is likely to be an intron with additional single base insertions or deletions in the flanking region on the right side. (D) Another coding insert. In this case the mismatch at the location of the capitalized C has to be introduced to get GT . . . AG ends at the insert. Note that 4 nucleotides to the left of the insert could be slid to the right of the insert and not require any mismatches.

eth et al. 1993) or that share an enhancer would also seem likely to be resistant to rearrangement. In part, the length of an alignment reflects the position on the chromosome. Eight of the 10 longest alignments, 15 of the 20 longest alignments, 36 of the 50 longest alignments, and 63 of the 100 longest alignments are found in the gene-rich clusters near the middle of chromosomes. The chromosome arms appear to be more sus-

Table 8. Profiles of Nucleotide Frequencies Adjacent to 5′ and 3′ Splice Sites

Base frequency profile near 5′ splice site																
A	a	32%	30%	33%	20%	32%	82%	4%	^	0.0%	0.0%	62%	51%	8%	23%	24%
	c	33%	17%	11%	40%	32%	5%	4%	^	0.0%	0.0%	0.0%	6%	14%	24%	12%
	g	11%	14%	17%	18%	24%	1%	88%	^	100%	0.0%	20%	18%	66%	4%	5%
	t	24%	39%	39%	22%	12%	12%	3%	^	0.0%	100%	17%	25%	13%	49%	59%
B	a	28%	29%	30%	34%	38%	63%	13%	^	0.0%	0.0%	62%	60%	12%	25%	29%
	c	24%	15%	17%	20%	20%	9%	5%	^	0.0%	0.0%	0.0%	4%	5%	8%	14%
	g	18%	28%	18%	13%	21%	10%	76%	^	100%	0.0%	15%	15%	63%	7%	9%
	t	29%	27%	34%	33%	21%	17%	6%	^	0.0%	100%	23%	21%	19%	60%	48%
C	a	28%	30%	35%	36%	40%	56%	18%	^	0.0%	0.0%	58%	66%	10%	19%	26%
	c	18%	18%	17%	18%	23%	14%	7%	^	0.0%	0.0%	2%	8%	3%	10%	10%
	g	23%	23%	17%	17%	20%	11%	60%	^	100%	0.0%	24%	9%	75%	9%	13%
	t	32%	28%	31%	28%	17%	18%	14%	^	0.0%	100%	16%	17%	11%	62%	51%
Base frequency profile near 3′ splice site																
D	a	19%	4%	1%	28%	0.0%	100%	0.0%	^	23%	30%	44%	22%	25%	17%	30%
	c	19%	5%	2%	12%	84%	0.0%	0.0%	^	25%	23%	24%	11%	44%	49%	20%
	g	9%	2%	0.0%	9%	0.0%	0.0%	100%	^	49%	11%	10%	18%	8%	4%	16%
	t	53%	88%	97%	52%	16%	0.0%	0.0%	^	3%	37%	23%	49%	24%	29%	33%
E	a	34%	7%	0.5%	11%	2%	100%	0.0%	^	29%	27%	38%	30%	31%	23%	36%
	c	7%	2%	0.5%	11%	80%	0.0%	0.0%	^	14%	20%	21%	24%	27%	32%	20%
	g	5%	4%	0.0%	8%	0.0%	0.0%	100%	^	46%	12%	16%	23%	17%	17%	23%
	t	55%	88%	99%	70%	18%	0.0%	0.0%	^	12%	41%	24%	23%	25%	28%	21%
F	a	28%	5%	1%	9%	3%	100%	0.0%	^	41%	30%	30%	28%	30%	28%	29%
	c	8%	3%	1%	16%	84%	0.0%	0.0%	^	16%	19%	23%	26%	26%	29%	26%
	g	6%	2%	0.3%	8%	0.1%	0.0%	100%	^	31%	16%	19%	22%	18%	19%	23%
	t	58%	90%	97%	67%	13%	0.0%	0.0%	^	13%	36%	28%	25%	26%	24%	22%

Profiles for the 93 introns unique to *C. briggsae* (A,D), the 170 introns unique to *C. elegans* (B,E), and the 28085 *C. elegans* introns that have canonical GT . . . AG ends (C,F). The caret (^) indicates the splice site. Note that the consensus of AG immediately before the 3′ splice site is much stronger for unique introns, particularly introns unique to *C. briggsae,* than it is for *C. elegans* introns as a whole.

Table 9. Percent Homology Between 3′ and 5′ Splice Sites

A	22	33	30	26	38	30	39	38	23	30	82	88	^	49	37	39	18	20	38	31	30	14	30	37	16
B	24	32	30	24	27	31	28	34	35	20	63	76	^	46	41	31	25	17	25	20	24	28	27	26	25
C	26	27	27	27	29	29	28	31	27	23	56	60	^	31	36	27	27	21	25	25	26	26	25	26	26

The table shows the average percent of matching bases for 12 nucleotides on either side of the splice site, which is shown by the caret (^). (A) Introns unique to *C. briggsae;* (B) introns unique to *C. elegans;* (C) all *C. elegans* introns.

ceptible to rearrangement. One cosmid, G46J06, aligns in nine widely separated pieces on chromosomes I, II, III, and IV. None of the G46J06 alignments are in the gene-rich cluster regions.

Conclusions

We developed a new algorithm for aligning genomic DNA. We used the algorithm to perform a large-scale alignment between *C. briggsae* and *C. elegans* genomic sequences. The alignment confirms that patterns of conservation can be useful in identifying regulatory regions and rarely expressed coding regions. The alignment, including *C. briggsae* homologs of 154 cloned *C. elegans* genes and many times this number of largely uncharacterized ORFs, can be viewed and searched in the Intronerator. The detailed alignment display can help identify conserved regulatory elements even inside exons. There is a bimodal distribution on the size of syntenic regions. Over 250 introns are present in one species but not the other, within the ~10% of the *C. briggsae* genome sequenced.

METHODS

Sources of Sequence Data

The genomic sequence data for *C. elegans* came from the *C. elegans* Sequencing Consortium (1998) downloaded from http://www.sanger.co.uk/Projects/C_elegans/chromosomes.shtml on August 4, 1999. The *C. briggsae* genomic sequence data came from the Washington University site ftp://genome.wustl.edu/pub/gsc1/sequence/st.louis/briggsae/finish/ on October 6, 1999. Genomic sequences from other organisms and all the mRNA sequences were obtained from GenBank (Benson et al. 1999) at http://www.ncbi.nlm.nih_.gov/Entrez/ on December 4, 1999.

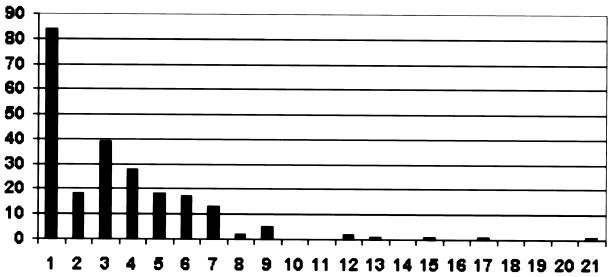

Figure 4 Synteny summary graph. The horizontal axis is how many pieces a *C. briggsae* clone needed to be broken into for maximal alignment with the *C. elegans* genome. The vertical axis is the number of clones broken into that many pieces. There were 229 clones, and the average clone size was 34,722 bases. Clones were broken up if the best aligning region of one 2000-base region of the *C. briggsae* clone aligned to a spot 50,000 bases or more away or on a different chromosome from where the previous 2000-base region of the *C. briggsae* clone aligned best. The 2000 base regions overlapped by 1000 bases. The average length of a broken-up region was 8553 bases. The bimodal distribution suggests that ~40% of the genome is resistant to rearrangements and the rest is not.

Cross-species DNA Alignment Algorithms

The large scale of the *C. briggsae* and *C. elegans* genomic alignment required a relatively fast algorithm. Yet the algorithm also needed to handle both small and large insertions and deletions and deal with the statistical challenges presented by the AT-richness of the *Caenorhabditis* genome. To cope with these somewhat conflicting concerns, we broke the alignment problem into three passes: a fast program that took overlapping 2000 base regions of *C. briggsae* and scanned the *C. elegans* genome for homologous regions, a program that aligned the homologous regions in a very detailed manner, and a program that stitched together the overlapping alignments.

The fastest DNA alignment algorithms, such as BLAST (Altschul et al. 1990), demand areas that match perfectly to serve as seeds for the alignment. Although such approaches work well for searching EST databases, they break down when comparing DNA sequences between species. Although coding regions of DNA are fairly well conserved in the first and second codon positions, the third, wobble, position is largely redundant in the genetic code and tends to diverge freely. To cope with this we developed an algorithm that worked on the general principles of gapped BLAST (Altschul et al. 1997) but that, rather than requiring six consecutive nucleotides to match perfectly to seed an alignment, required six nucleotides spread out in the pattern XXoXXoXX (where the X's must match but the o's need not) to serve as the seed. We call this algorithm WABA. There are three passes to the WABA algorithm.

The first pass of WABA works in the following fashion: The *C. elegans* DNA (the target sequence) is packed so that each 16-bit word contains 8 nucleotides. This serves two purposes—to reduce the amount of RAM required to store the genome and to allow multiple nucleotides to be compared in a single computer operation. For each section of the *C. briggsae* sequence (the query sequence) we do the following: An array that contains an entry for all possible 8-nucleotide sequences (8-mers) is allocated. The query sequence is then scanned, and each 8-mer is checked and possibly added to the array. Note that these 8-mers in the query sequence overlap each other by 7 bases. Eight-mers are rejected if they contain ambiguous nucleotides such as N or if they could match degenerate repeating DNA sequences of period 4 or less. (This eliminates 8-mers such as AAAAAAAA, ACACACAC, ACGACGAC, and ACGTACGT.) The 8-mers are packed into a 16-bit

word, which is then binary logical anded with the wobble mask (hexadecimal 0xF3CF) so that the third and sixth nucleotides are ignored. The packed, masked 8-mer is then added to the corresponding entry in the array that stores the position of the 8-mer in the query sequence as well. Then, the packed genome is scanned. Each packed word of the genome is "anded" with the wobble mask, and the result is used to index the array. If the corresponding array entry is nonempty, then the current genome position and the position of the corresponding query 8-mer are noted in a hit list. Because the hit list grows quite quickly, it is periodically scanned and nonpromising hits are eliminated. If a hit occurs within 1000 bases on the target sequence of another hit and the positions of the hits in both query and target indicate that they could be part of a homologous region lacking inserts (which is simply checked by subtracting from query offsets of both hits to get a "diagonal" offset and then checking that the diagonal offsets are identical), it is considered promising. In this way, WABA scans the entire genome and assembles a list of hits. The hits are then clumped together. A clump starts with a single hit. Each hit within 48 nucleotides in both target and query and with the same diagonal offset is added to a clump. As hits are added to a clump, the boundaries of the clump are extended to include the new hits. The clumps are then scored by the square of the number of hits in the clump, and the clumps scoring <25% of the highest score are eliminated. Because the distribution of scores is highly peaked, this saves processing in later steps in the common scenario where there are one or a few areas of strong homology and many areas of weak homology. The remaining clumps are scored for the best local alignment without inserts between the full query sequence and the corresponding part of the target sequence. Because the worm genome is AT rich, AT matches are scored as 2, whereas GC matches are scored as 3. Mismatches are scored as 4. The bottom scoring 25% of the alignments without inserts are dropped, and the remaining alignments are stored for the next pass.

Though the first pass is quite fast, it has no provision for insertions and deletions. The first pass merely identifies regions of the target homologous to the 2000-base query sequence. A 5000-base window of the target sequence centered around the first pass aligment is then aligned with a slower algorithm that can cope with insertions and deletions. The Smith–Waterman algorithm (Smith and Waterman 1981) and extensions of it to allow affine gap scores (Pearson and Miller 1992) would serve this purpose reasonably well. However, even affine gap penalties tend to penalize long gaps excessively in genomic and cDNA alignments. Furthermore, we wanted our slow, detailed aligner to be at least as wobble-aware as the fast front end. HMMs have proven to be very useful in predicting gene structure (Kulp et al. 1996). Pair HMMs of three (hidden) states can provide a conceptually simple and effective implementation of alignment with affine gap costs (Durbin et al. 1998). This motivated us to design a seven state pair HMM to do the detailed alignment (Fig. 5). One state captures long inserts in the target sequence, one state captures long inserts in the query sequence, one state captures highly conserved regions (~90% base similarity), one state captures lightly conserved regions (~50% base similarity), and three states capture coding regions. As the HMM matches a nucleotide pair in one coding state, it shifts to the next coding state. Match and mismatch scores for the first two coding states are very similar to match and mismatch scores for the highly conserved regions. Match and mismatch scores

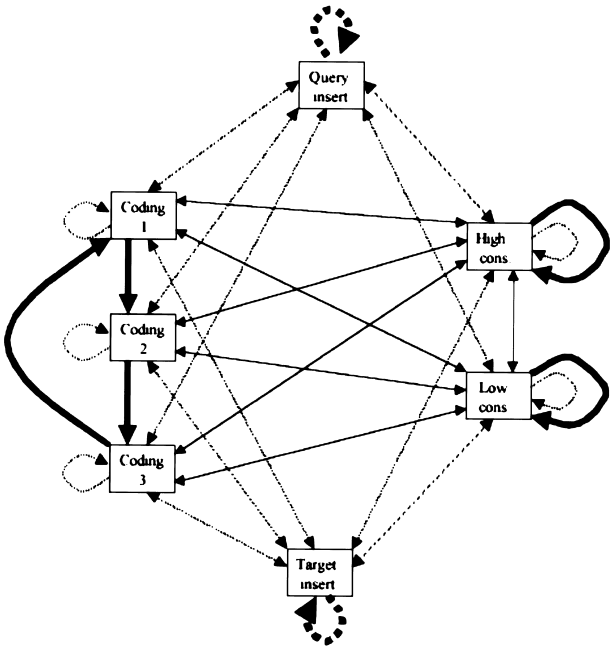

Figure 5 Finite state machine depiction of seven state aligner. This is a pairwise HMM. Transitions between states are associated either with aligning a pair of nucleotides (solid arrows) or aligning a single nucleotide with a gap (broken line arrows). The most likely transition out of each state is shown by a thick arrow. Thin arrows represent possible, but unlikely, transitions. In coding regions, the most likely path will typically cycle between coding states 1, 2, and 3. In strongly conserved regions, the most likely path will typically stay in the high conservation state. In weakly conserved regions, the most likely path will typically stay in the low conservation state. Short inserts follow the broken line arrows in the coding and conserved states. Long inserts transition into the query or target insert states.

for the third coding state are lower and favor mismatches between C and T or G and A over other mismatches. The heavily and lightly conserved and coding states all allow inserts within the state with a gap penalty that is approximately equivalent to eight mismatches. To transition into a long gap state invokes a penalty equivalent to roughly 12 mismatches. To continue a gap in a long gap state costs a penalty that is only about one-fifth of a mismatch cost. The result is an algorithm that penalizes 1-base gaps by 8, 2-base gaps by 12, and longer gaps by 12 plus one-fifth of their length. A 60-bp gap is only penalized twice as much as a 2-bp gap. This is a fair approximation of the distribution of gap sizes in genomic alignments (data not shown) that is not as computationally expensive as more elaborate gap penalty schemes (Gotoh 1990). The matches are also somewhat more sensitive than the usual practice of treating all nucleotide matches equivalently. Because the worm genome is 61% A/T, we weighed G/C matches more heavily than A/T matches. The various match, mismatch, and gap scores were chosen first by an educated guess of the equivalent probabilities. The resulting HMM was used to align a substantial fraction of the available data. The distributions of matches, mismatches, and gaps in this first alignment were used to refine the scores for a second alignment over the entire data set. This provided one step of parameter optimization by expectation maximization (Durbin et al. 1998). The result is a robust and sensitive align-

ment algorithm. It also produces predictions as to whether a region is coding or not and a score that can be used to compare the significance of different alignments. We empirically set the scoring threshold for an alignment to pass onto the final pass of WABA to a number that just excluded a large number of ≈40-bp alignments that match primarily in poly-A and poly-T regions. This is necessary because the worm genome is very rich in these regions. The oligonucleotide AAAA occurs 7.5 times as often as would be expected from a random distribution of nucleotides.

The pairwise HMM that is the second pass of WABA is implemented in optimized C code. Nonetheless, because the run time increases with the product of the length of the two sequences that it is aligning, it is not efficient to use it for long alignments. This is the primary motivation for breaking the query sequence into 2000-bp blocks that overlap each other by 1000 bp. This necessitates a third pass to scan the resulting short alignments and, when possible, merge them into longer alignments. This last pass works by looking for alignments that overlap in both query and target. Overlapping areas are scanned for regions of at least 15 nucleotides where the alignments are identical and are joined if such a region is found. After this joining, the alignment is complete.

ACKNOWLEDGMENTS

We thank David Haussler for his sage advice on HMMs, David Kulp for making Genie predictions for the Intronerator, and members of the Zahler, Jin, and Chisholm labs and Manny Ares for helpful discussions. Thanks to David Haussler, David Hoffman, and Jane Silverthorne for critical reading of this manuscript. We are indebted to all the various sequencing projects for providing sequence data. This work was supported by grant 1R01GM52848 from the National Institutes of Health to A.M.Z. and a training grant from the University of California Biotechnology Program.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

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

Ayala, F.J., Barrio, E., and Kwiatowski, J. 1996. Molecular clock or erratic evolution? A tale of two genes. *Proc. Natl. Acad. Sci.* **93:** 11729–11734.

Bailey, T.L. and Elkan, C. 1995. Unsupervised learning of multiple motifs in biopolymers using EM. *Mach. Learn.* **21:** 51–80.

Baldwin, J.G., Frisse, L.M., Vida, J.T., Eddleman, C.D., and Thomas, W.K. 1997. An evolutionary framework for the study of developmental evolution in a set of nematodes related to *Caenorhabditis elegans*. *Mol. Phylogenet. Evol.* **8:** 249–259.

Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, B.F., Rapp, B.A., and Wheeler, D.L. 1999. GenBank. *Nucleic Acids Res.* **27:** 12–17.

Blaxter, M. 1998. *Caenorhabditis elegans* is a nematode. *Science* **282:** 2041–2046.

Blaxter, M.L., De Ley, P., Garey, J.R., Liu, L.X., Scheldeman, P., Vierstraete, A., Vanfleteren, J.R., Mackey, L.Y., Dorris, M., Frisse, L.M., et al. 1998. A molecular evolutionary framework for the phylum Nematoda. *Nature* **392:** 71–75.

Boeddrich, A., Burgtorf, C., Roest Crollius, H. Hennig, S., Bernot, A., Clark, M., Reinhardt, R., Lehrach, H., and Francis, F. 1999. Analysis of the spermine synthase gene region in *Fugu rubripes*, *Tetraodon fluviatilis*, and *Danio rerio*. *Genomics* **57:** 164–168.

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

Durbin, R.E., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.

Eastman, C., Horvitz, H.R., and Jin, Y. 1999. Coordinated transcriptional regulation of the *unc-25* glutamic acid decarboxylase and the *unc-47* GABA vesicular transporter by the *Caenorhabditis elegans* UNC-30 homeodomain protein. *J. Neurosci.* **19:** 6225–6234.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95:** 14863–14868.

Endrizzi, M., Huang, S., Scharf, J.M., Kelter, A.R., Wirth, B., Kunkel, L.M., Miller, W., and Dietrich, W.F. 1999. Comparative sequence analysis of the mouse and human Lgn1/SMA interval. *Genomics* **60:** 137–151.

Gotoh, O. 1990. Optimal sequence alignment allowing for long gaps. *Bull. Math. Biol.* **52:** 359–373.

Hansen, D. and Pilgrim, D. 1998. Molecular evolution of a sex determination protein. FEM-2 (pp2c) in Caenorhabditis. *Genetics* **149:** 1353–1362.

Hardison, R.C., Oeltjen, J., and Miller, W. 1997. Long human–mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7:** 959–966.

Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9:** 815–824.

Jeggo, P.A. 1998. DNA breakage and repair. *Adv. Genet.* **38:** 185–218.

Kent, W.J. and Zahler, A.M. 2000. The Intronerator: Exploring introns and alternative splicing in *C. elegans*. *Nucleic Acids Res.* **28:** 91–93.

Kulp, D., Haussler, D., Reese, M.G., and Eeckman, F.H. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. *ISMB* **4:** 134–142.

———. 1997. Integrating database homology in a probabilistic gene structure model. *Pac. Symp. Biocomput.* 232–244.

Makalowski, W. and Boguski, M.S. 1998. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci.* **95:** 9407–9412.

Makalowski, W., Zhang, J., and Boguski, M.S. 1996. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.* **6:** 846–857.

McIntire, S.L., Reimer, R.J., Schuske, K., Edwards, R.H., and Jorgensen, E.M. 1997. Identification and characterization of the vesicular GABA transporter. *Nature* **389:** 870–876.

Nagel, R.J., Lancaster, A.M., and Zahler, A.M. 1998. Specific binding of an exonic splicing enhancer by the pre-mRNA splicing factor SRp55. *RNA* **4:** 11–23.

O'Brien, S.J., Menotti-Raymond, M., Murphy, W.J., Nash, W.G., Wienberg, J., Stanyon, R., Copeland, N.G., Jenkins, N.A., Womack, J.E., and Marshall Graves, J.A. 1999. The promise of comparative genomics in mammals. *Science* **286:** 458–462; 479–481.

Pearson, W.R. and Miller, W. 1992. Dynamic programming algorithms for biological sequence comparison. *Methods Enzymol.* **210:** 575–601.

Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. 1997. Improved splice site detection in Genie. *J. Comput. Biol.* **4:** 311–323.

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

Shabalina, S.A. and Kondrashov, A.S. 1999. Pattern of selective

constraint in *C. elegans* and *C. briggsae* genomes. *Genet. Res.* **74:** 23–30.
Sibley, M.H., Johnson, J.J., Mello, C.C., and Kramer, J.M. 1993. Genetic identification, sequence, and alternative splicing of the *Caenorhabditis elegans* alpha 2(IV) collagen gene. *J. Cell Biol.* **123:** 255–264.
Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9:** 657–663.
Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.
Spieth, J., Brooke, G., Kuersten, S., Lea, K., and Blumenthal, T. 1993. Operons in *C. elegans*: Polycistronic mRNA precursors are processed by *trans*-splicing of SL2 to downstream coding regions. *Cell* **73:** 521–532.
Thacker, C., Marra, M.A., Jones, A., Baillie, D.L., and Rose, A.M. 1999. Functional genomics in *Caenorhabditis elegans*: An approach involving comparisons of sequences from related nematodes. *Genome Res.* **9:** 348–359.
Voronov, D.A., Panchin, Y.V., and Spiridonov, S.E. 1998. Nematode phylogeny and embryology. *Nature* **395:** 28.
Wilson, W.A., Harrington, S.E., Woodman, W.L., Lee, M., Sorrells, M.E., and McCouch, S.R. 1999. Inferences on the genome structure of progenitor maize through comparative analysis of rice, maize and the domesticated panicoids. *Genetics* **153:** 453–473.

Received January 27, 2000; accepted in revised form June 2, 2000.