Automated annotation of *Caenorhabditis* mitochondrial genomes and phylogenetic analysis

Jessica L. Campbell[1], Larry J. Wilhelm[2], Dee R. Denver[2]

[1]Department of Bioresource Research, Oregon State University, Corvallis, Oregon 97331.

[2]Department of Zoology, Oregon State University, Corvallis, Oregon 97331.

Abstract

Motivation

Relatively little is known about the evolution of mitochondrial genomes between *Caenorhabditis* species despite decades of research. Both the worldwide search for new nematode species and the effort to increase sequencing speeds require automated tools to quickly characterize and categorize large amounts of data.

Results

We developed an automated annotation tool for nematode mitochondrial genome sequences that characterizes individual genes in addition to pseudogenic and intergenic regions. The automated annotation tool utilizes the ClustalW multiple sequence alignment program to produce gene alignments of input genomic sequence versus literature *C. elegans* mitochondrial genes.

A phylogenetic analysis was conducted by using RAxML 7.2.6 to create a maximum-likelihood tree with bootstrap analysis for 84 mitochondrial genomic sequences representing 23 *Caenorhabditis* species and 1 outgroup.

Conclusion

The automated annotation tool for *Caenorhabditis* mitochondrial genome sequences is fast and effective. The phylogram resulting from phylogenetic analysis provided preliminary insights into the evolutionary relationships among several *Caenorhabditis* species' mitochondrial genomes.

1 INTRODUCTION

Nematodes are one of the most abundant and most diverse animals on Earth, with an estimated 0.5 to 100 million species (Dorris *et al*., 1999). One such nematode, the soil-dwelling *Caenorhabditis elegans*, was the first animal to have its genome sequenced and has since served as a useful model organism (*C. elegans* Sequencing Consortium, 1998). The *C. elegans* genome is not ideal for comparative genomic analysis because closely related species have yet to be discovered, resulting in a global search for a "sister species". This search has led to the discovery of several species of nematode classified under the *Caenorhabditis* genus, which require extensive phylogenetic analysis to determine their relationship within the *Caenorhabditis* genus and to *C. elegans* specifically.

The mitochondrial genome is vital for animal metabolism and physiology. The typical animal mitochondrial genome is easily comparable to *C. elegans*' mitochondrial genome. The *C. elegans* mitochondrial genome is 13,794 nucleotides long and includes 36 genes: 22 transfer RNAs, 12 protein-coding genes and 2 ribosomal RNAs (Okimoto *et al*., 1992).

Although some nematode groups share a common order of mitochondrial genes, there is extensive variation in mitochondrial gene arrangements throughout the phylum Nematoda. For example, the human parasite *Strongyloides stercoralis* (Hu et al., 2003) and the human pinworm *Enterobius vermicularis* (Kang *et al*., 2009) have mitochondrial genomes with an extremely rearranged gene order. Some nematode mitochondrial DNA encodes duplicated gene regions and/or pseudogenic regions (Howe and Denver: 2008; Hyman *et al*., 1998; Tang and Hyman, 2007). Plant-parasitic nematodes in the genus *Globodera* show the extremely unusual phenomenon of multi-chromosomal mitochondrial DNA (Gibson *et al*., 2007).

Although mitochondrial genomes have been heavily studied for decades, knowledge of the forces that shape mitochondrial genomic evolution and the variation between animal lineages is

limited. Determining the underpinnings of mitochondrial genome evolution is a goal of many laboratories.

The development of an automated tool to annotate mitochondrial genes is essential to efficiently identify the patterns of mitochondrial genome variations either across several generations and/or between species. With the increasing rate of high-throughput sequencing, tools need to be developed to analyze the massive amounts of sequence data en masse.

Information regarding the evolutionary history of *Caenorhabditis* mitochondrial genes and evolutionary relationships between *Caenorhabditis* species is extremely limited. Through phylogenetic analyses, we can begin to describe the evolutionary relatedness between species in the *Caenorhabditis* genus.

2 MATERIALS AND METHODS

2.1 Mitochondrial genome annotation

The annotation tool utilized the command line ClustalW multiple sequence alignment program (Larkin, et al, 2007) to determine homologous genes by aligning literature *C. elegans* mitochondrial genes against user-submitted input sequences. ClustalW was chosen for use within the automated annotation tool because of its novel position-scoring and weighting algorithms as well as its speed and low resource use.

The annotation tool itself was written in Perl and utilized several BioPerl modules and NCBI's database formatting tools.

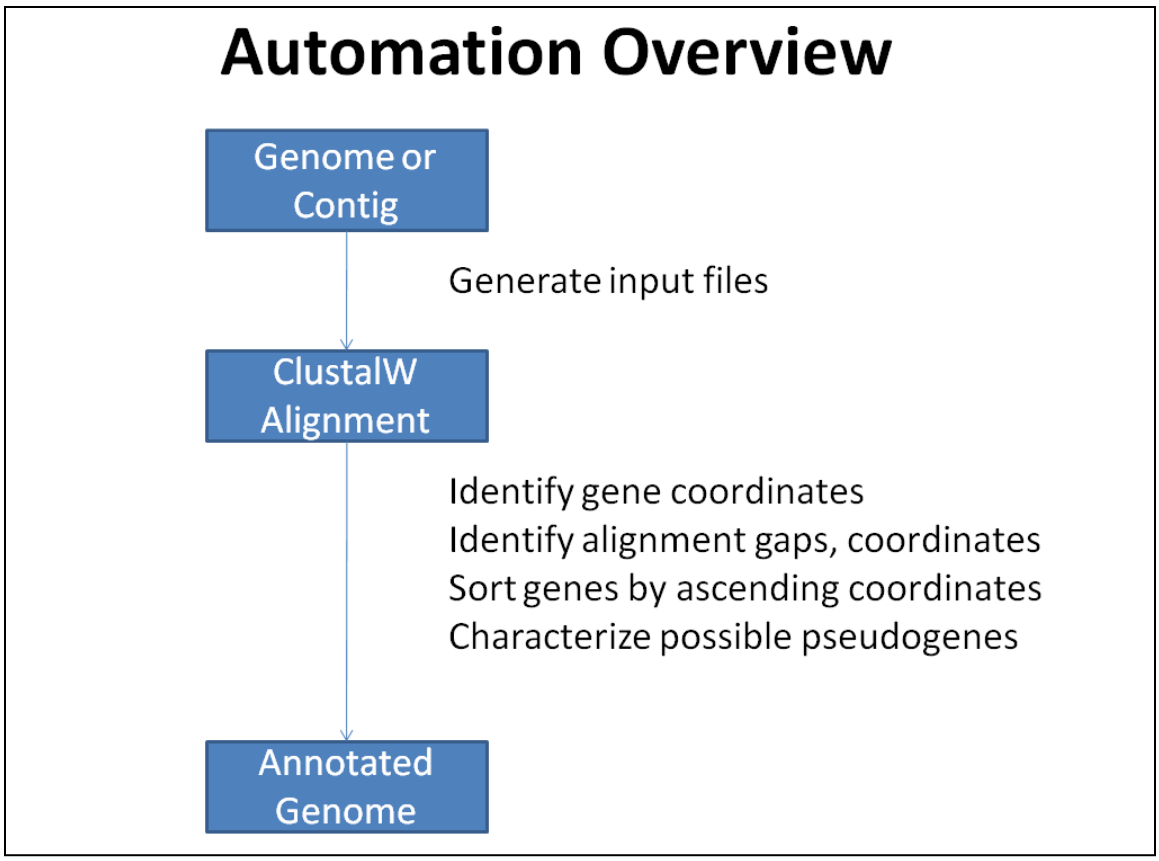

Figure 1. Stepwise overview of the automated annotation pipeline from initial input of contigs or genomes to completed annotation.

2.1.1 Identification of gene boundaries

The automated annotation tool is ran from the command line with the required input of a FASTA file containing a reference genome or contig of interest. The *C. elegans* reference mitochondrial genes obtained from GenBank were hardcoded locally. Using the BioPerl SeqIO module (**Appendix 5.1**), the tool created FASTA files containing the *C. elegans* reference mitochondrial gene sequence and the user-submitted genomic sequence. These newly generated FASTA files served as the input for ClustalW DNA alignment.

Running ClustalW DNA alignments using the newly generated FASTA files created alignment files documenting ClustalW's results. The alignment files were used to compare the input genomic sequence against the literature *C. elegans* mitochondrial genes sequences to identify gene boundaries within the input sequence and alignment gaps (**Appendix 5.2**). The newly identified

genetic boundaries were translated into coordinates relative to the genomic sequence; e.g., 1 corresponded to the first nucleotide in the provided genomic sequence. Alignment gaps occurred when a reference *C. elegans* gene best matched a portion of the input sequence, but only partially aligned due to significant differences in nucleotide sequence. These alignment gaps were noted for future analyses, as they may be indicative of evolutionary changes between the species of interest and the reference *C. elegans* genome.

The mitochondrial genes were sorted by ascending gene coordinates through the use of an anonymous multidimensional hash. The hash generated a database that represents a rudimentary genetic map of the mitochondrial genome. The database contained the alignment name (gene name_v_sequence name), the gene coordinates within the input sequence, the ClustalW alignment score, and the alignment gap coordinates.

2.1.2 Pseudogene characterization

Intergenic regions were analyzed independently to identify and characterize pseudogenes (**Figure 2**). Intergenic regions were typically very small due to the compact nature of mitochondrial genomes. Because pseudogenes tend to arise from retrotransposition, the integration of a recently transcribed mRNA back into the genome, pseudogenes tend to be larger than intergenic regions and can be traced back to their genetic origins.

To account for this, the tool excised intergenic sequences greater than a user-defined threshold from the initial input sequence. Given no user input, the default threshold was two hundred nucleotides. A FASTA file was generated containing the excised intergenic sequence and the original input sequence without the excised intergenic region. This FASTA file was used to perform a separate ClustalW DNA alignment. The alignment file generated by ClustalW was used to identify gene coordinates and alignment gaps using the algorithm described in Section 2.1.

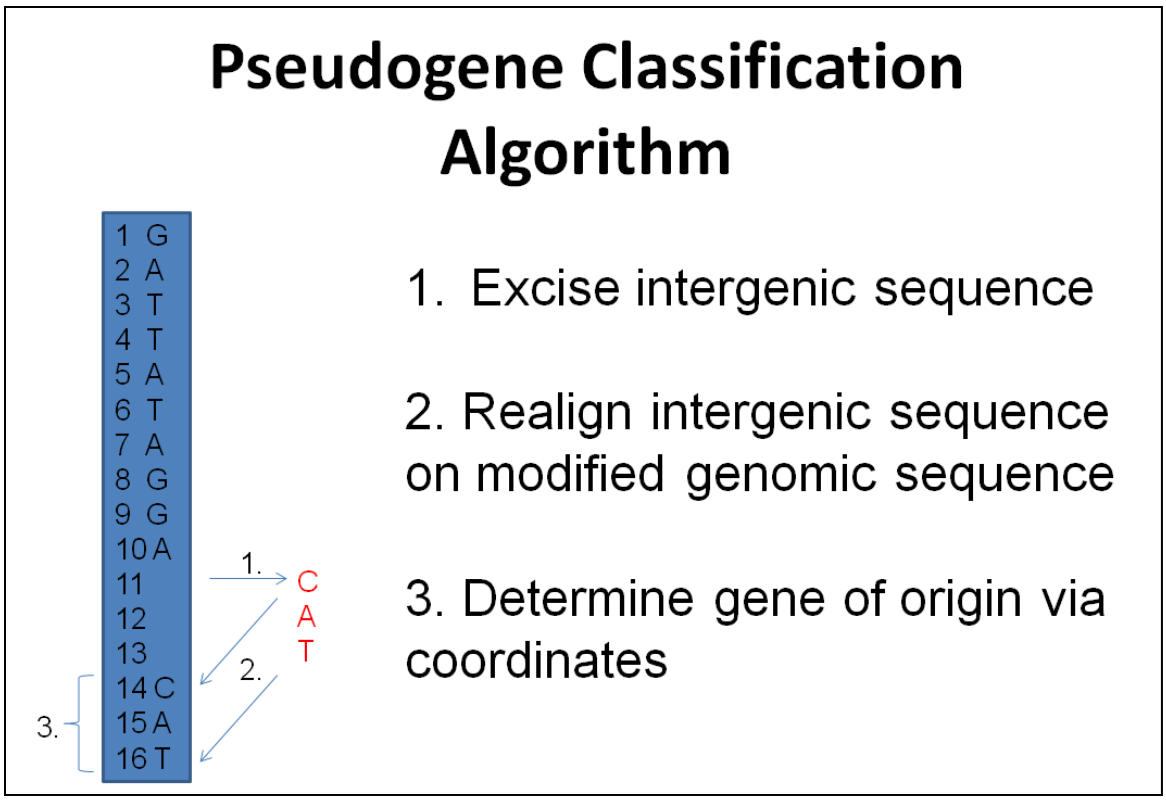

Figure 2. Stepwise example of the pseudogene classification algorithm.

The resulting ClustalW alignment score, the percentage of the excised intergenic sequence that aligned against the input sequence minus the excised intergenic sequence, determined whether the pseudogene characterization proceeds. If the ClustalW alignment score was below the user-defined threshold, or below 50 if no user input is given, the characterization halted.

If the ClustalW alignment score was above the threshold, the coordinates of where the intergenic region aligned were compared to the previously generated database of mitochondrial gene coordinates in the original input sequence to determine the gene of origin. The final output was the gene of origin of the pseudogenic region and its ClustalW alignment score.

2.2 Phylogenetic analysis

A phylogenetic analysis of 84 mitochondrial genomic sequences from 23 *Caenorhabditis* species and *Pristionchus pacificus* was conducted (**Table 1**).

Species	**Isolate**	**Origin**
Caenorhabditis briggsae	BW287	Beijing, China

Caenorhabditis briggsae	ED3032	Taipei, Taiwan
Caenorhabditis briggsae	ED3033	Taipei, Taiwan
Caenorhabditis briggsae	ED3034	Taipei, Taiwan
Caenorhabditis briggsae	ED3035	Taipei, Taiwan
Caenorhabditis briggsae	ED3036	Taipei, Taiwan
Caenorhabditis briggsae	ED3037	Taipei, Taiwan
Caenorhabditis briggsae	ED3083	Johannesburg, South Africa
Caenorhabditis briggsae	ED3092	Nairobi, Kenya
Caenorhabditis briggsae	ED3101	Nairobi, Kenya
Caenorhabditis briggsae	EG4181	Utah, USA
Caenorhabditis briggsae	EG4207A	Utah, USA
Caenorhabditis briggsae	HK104	Okayama, Japan
Caenorhabditis briggsae	HK105	Sendai, Japan
Caenorhabditis briggsae	JU403	Hermanville, France
Caenorhabditis briggsae	JU439	Reykjavic, Iceland
Caenorhabditis briggsae	JU516	Marsas, France
Caenorhabditis briggsae	JU725	Chengyang, China
Caenorhabditis briggsae	JU726	Tangshuo, China
Caenorhabditis briggsae	JU793	Frechendets, France
Caenorhabditis briggsae	JU1424	Ba Be Lake, Vietnam
Caenorhabditis briggsae	OR24	
Caenorhabditis briggsae	PB800	
Caenorhabditis briggsae	PB826	
Caenorhabditis briggsae	VT847	
Caenorhabditis elegans	AB1	Adelaide, Australia
Caenorhabditis elegans	AB2	Adelaide, Australia
Caenorhabditis elegans	CB4852	England
Caenorhabditis elegans	CB4853	Altadena, California, USA
Caenorhabditis elegans	CB4854	Altadena, California, USA
Caenorhabditis elegans	CB4855	Palo Alto, California, USA
Caenorhabditis elegans	CB4856	Hawaii, USA
Caenorhabditis elegans	CB4857	Claremont, California, USA
Caenorhabditis elegans	CB4858	Pasadena, California, USA
Caenorhabditis elegans	DL0200	St. Joseph, Missouri, USA
Caenorhabditis elegans	JU258	
Caenorhabditis elegans	KR314	Vancouver, Canada
Caenorhabditis elegans	N2	Bristol, England
Caenorhabditis elegans	PB303	Fort Bragg, California, USA
Caenorhabditis elegans	PB306	Glouchester, Massachusetts, USA
Caenorhabditis elegans	PS2025	
Caenorhabditis elegans	RW7000	Bergerac, France
Caenorhabditis elegans	TR403	Madison, Wisconsin, USA
Caenorhabditis brenneri	JU1324	Poovar, Kerala, India
Caenorhabditis brenneri	JU1379	La Reunion, France
Caenorhabditis brenneri	JU1398	Medelin, Columbia
Caenorhabditis brenneri	JU1816	Cacao, French Guiana
Caenorhabditis brenneri	LKC28	Costa Rica
Caenorhabditis brenneri	SB280	Guadeloupe, France
Caenorhabditis remanei	DL271	Madison, Wisconsin, USA

Caenorhabditis remanei	JU825	Obernai, Bas-Rhin, France
Caenorhabditis remanei	JU1082	Okazaki, Japan
Caenorhabditis japonica	DF5081	Takeo, Japan
Caenorhabditis sp. 1	SB341	Berlin, Germany
Caenorhabditis sp. 3	RGD1	Homestead, Florida, USA
Caenorhabditis sp. 3	PS1010	Dade County, Florida, USA
Caenorhabditis sp. 5	JU737	
Caenorhabditis sp. 5	JU1423	Ba Be Lake, Vietnam
Caenorhabditis sp. 5	SB378	Guangzhou, China
Caenorhabditis sp. 6	EG4788	Amares, Portugal
Caenorhabditis sp. 7	JU1593	Shonga, Nigeria
Caenorhabditis sp. 8	QX1182	New Jersey, USA
Caenorhabditis sp. 9	EG5268	Congo, Africa
Caenorhabditis sp. 9	JU1325	Trivandrum, Kerala, India
Caenorhabditis sp. 10	JU1328	Kanjirapally, Kerala, India
Caenorhabditis sp. 10	JU1330	Kanjirapally, Kerala, India
Caenorhabditis sp. 10	JU1333	Periyar, Kerala, India
Caenorhabditis sp. 11	JU1373	La Reunion, France
Caenorhabditis sp. 11	JU1428	Nouragues Forest, French Guiana
Caenorhabditis sp. 11	JU1630	Santo Antao Island, Cape Verde
Caenorhabditis sp. 11	JU1639	Santo Antao Island, Cape Verde
Caenorhabditis sp. 11	JU1640	Santiago Island, Cape Verde
Caenorhabditis sp. 11	JU1818	Kaw Mountain, French Guiana
Caenorhabditis sp. 12	JU1426	Nouragues Forest, French Guiana
Caenorhabditis sp. 12	JU1427	Nouragues Forest, French Guiana
Caenorhabditis sp. 13	JU1528	Orsay Orchard, France
Caenorhabditis sp. 14	EG5716	Moorea, Tahiti
Caenorhabditis sp. 15	QG122	Kaui, Hawaii, USA
Caenorhabditis sp. 16	JU1873	Sanda Center, Bali, Indonesia
Caenorhabditis sp. 17	JU1825	Nouragues Forest, French Guiana
Caenorhabditis sp. 17	NIC59	Kourou, French Guiana
Caenorhabditis sp. 18	JU1857	Nouragues Forest, French Guiana
Caenorhabditis sp. 19	EG6142	Puerto Rico
Pristionchus pacificus		GenBank Accession ABKE00000000

Table 1. Species, isolates, and origins used for phylogenetic analysis. The sequences were sequenced by amplifying the mitochondria as two overlapping regions via long PCR. The amplicons were then used as input for Illumina high-throughput sample preparation and De Novo assembly was completed with SCRAPE.

The 84 mitochondrial genomic sequences were aligned using the MUSCLE multiple sequence alignment program. Alignment gaps longer than five nucleotides long were manually removed in MEGA 4.1.

The data-containing FASTA file was transformed to a PHYLIP file using ReadSeq and an improvised script to modify mitochondrial genome names to prevent sequence name clashes in following scripts.

RAxML (Randomized Axelerated Maximum Likelihood) 7.2.6 was used to conduct a rapid bootstrap analysis (Stamatakis *et al*.) over 1000 replicates and search for the best-scoring maximum-likelihood using a General Time Reversible model of nucleotide substitution and the I' model of hereogeneity (Stamatakis, 2006).

Bootstrapping tests the reliability of a tree based on substitution matrices. From each sequence, *n* nucleotides were randomly chosen with replacements, giving rise to *m* rows of *n* columns each. A maximum-likelihood tree was built with these newly created matrices. The topology of this new maximum-likelihood tree was compared to the topology of the original tree. Each branch was assigned a value of 0 if it differed from the original tree, or 1 if was the same. The resampling sites and tree reconstruction procedure was repeated 1000 times. The percentage of times each interior branch is given a value of 1 was recorded.

3 RESULTS

3.1 Annotation tool

3.1.1 Positive control: *C. elegans* reference genome

The annotation tool correctly identified the literature gene coordinates of the 34 *C. elegans* mitochondrial genes on the DL0200 *C. elegans* reference genome (**Table 2**). *C. elegans* gene coordinates were independently verified through multiple alignments using Molecular Evolutionary Genetics Analysis (MEGA).

The ClustalW alignment scores represent 98-100% alignment (Table 2), which is to be expected using literature *C. elegans* mitochondrial genes.

<Alignment>	<Coordinates>	<Score>	<Gap Coordinates>
tRNA_1_v_DL0200.aln	1-55	100	
tRNA_2_v_DL0200.aln	58-112	100	
CDS_3_v_DL0200.aln	113-547	98	
CDS_4_v_DL0200.aln	549-782	99	
tRNA_5_v_DL0200.aln	785-841	100	
tRNA_6_v_DL0200.aln	842-897	100	
tRNA_7_v_DL0200.aln	1595-1647	100	
tRNA_8_v_DL0200.aln	1648-1703	100	
tRNA_9_v_DL0200.aln	1707-1762	100	
CDS_10_v_DL0200.aln	1763-2638	98	
CDS_11_v_DL0200.aln	2634-3233	98	
tRNA_12_v_DL0200.aln	3240-3302	100	
tRNA_13_v_DL0200.aln	3303-3357	100	
tRNA_14_v_DL0200.aln	3358-3413	100	
CDS_15_v_DL0200.aln	3414-4262	98	
tRNA_16_v_DL0200.aln	4265-4325	100	
tRNA_17_v_DL0200.aln	4326-4380	100	
tRNA_18_v_DL0200.aln	4381-4435	100	
tRNA_19_v_DL0200.aln	4443-4499	100	
CDS_20_v_DL0200.aln	4500-5612	98	
tRNA_21_v_DL0200.aln	5617-5673	100	
CDS_22_v_DL0200.aln	5674-6441	99	
tRNA_23_v_DL0200.aln	6446-6501	100	
CDS_24_v_DL0200.aln	6502-7731	99	
CDS_25_v_DL0200.aln	7842-9419	98	
tRNA_26_v_DL0200.aln	9419-9474	100	
tRNA_27_v_DL0200.aln	9475-9534	100	
tRNA_28_v_DL0200.aln	9535-9589	100	
tRNA_29_v_DL0200.aln	9590-9645	98	
CDS_30_v_DL0200.aln	9646-10341	98	
tRNA_31_v_DL0200.aln	10345-10399	100	
CDS_32_v_DL0200.aln	11353-11688	98	
CDS_33_v_DL0200.aln	11688-13271	98	
tRNA_34_v_DL0200.aln	13272-13325	100	

Table 2. Genetic coordinate database generated from *C. elegans* reference genome.

3.1.2 Negative control: *C. briggsae* reference genome

The correct gene coordinates for the JU1424 *C. briggsae* reference genome were identified (**Table 3**), as independently verified through multiple alignments using MEGA.

The *C. briggsae* mitochondrial genome contained two pseudogenes, ψnad5-1 and ψnad5-2, both originating from the NADH dehydrogenase 5 protein-coding gene referred to as nad5 (Howe and Denver: 2008). The annotation tool correctly identified the protein-coding, or nad5, gene in the JU1424 *C. briggsae* reference genome without implicating either of the pseudogenes (**Table 3:** CDS_32).

<Alignment>	<Coordinates>	<Score>	<Gap Coordinates>
tRNA_1_v_JU1424.aln	1-55	96	
tRNA_2_v_JU1424.aln	58-112	100	
CDS_3_v_JU1424.aln	113-547	82	
CDS_4_v_JU1424.aln	550-783	89	
tRNA_5_v_JU1424.aln	783-839	96	
tRNA_6_v_JU1424.aln	840-895	98	
tRNA_7_v_JU1424.aln	1592-1644	96	
tRNA_8_v_JU1424.aln	1645-1701	96	1661-1661
tRNA_9_v_JU1424.aln	1705-1762	94	1727-1727, 1751-1751
CDS_10_v_JU1424.aln	1763-2638	85	
CDS_11_v_JU1424.aln	2634-3233	87	
tRNA_12_v_JU1424.aln	3239-3301	96	
tRNA_13_v_JU1424.aln	3302-3356	96	
tRNA_14_v_JU1424.aln	3357-3412	91	
CDS_15_v_JU1424.aln	3413-4261	85	
tRNA_16_v_JU1424.aln	4264-4325	96	4280-4280
tRNA_17_v_JU1424.aln	4326-4381	78	4375-4375
tRNA_18_v_JU1424.aln	4384-4438	89	
tRNA_19_v_JU1424.aln	4661-4717	91	
CDS_20_v_JU1424.aln	4718-5830	86	
tRNA_21_v_JU1424.aln	5835-5891	91	
CDS_22_v_JU1424.aln	5891-6658	86	
tRNA_23_v_JU1424.aln	6663-6718	94	
CDS_24_v_JU1424.aln	6719-7948	85	
CDS_25_v_JU1424.aln	8059-9636	87	
tRNA_26_v_JU1424.aln	9636-9691	91	
tRNA_27_v_JU1424.aln	9692-9751	95	
tRNA_28_v_JU1424.aln	9752-9806	98	
tRNA_29_v_JU1424.aln	9808-9863	89	
CDS_30_v_JU1424.aln	9864-10559	89	
tRNA_31_v_JU1424.aln	10563-10618	94	10608-10608
CDS_32_v_JU1424.aln	11573-11908	81	
CDS_33_v_JU1424.aln	12251-13834	86	
tRNA_34_v_JU1424.aln	13837-13890	85	

Table 3. Genetic coordinate and alignment gap database generated from *C. elegans* reference genome.

The annotation tool correctly identified the intergenic regions that contained pseudogenes and characterized their gene of origin using a ClustalW alignment score threshold of 80. The tool identified ψnad5-1 between tRNA_6 and tRNA_7 [4] and correctly identified nad5 (CDS_32) as the gene of origin (**Table 4**). The ClustalW alignment score was 96. The tool also identified ψnad5-2 between tRNA_31 and the nad5 gene (CDS_32) [4] and identified nad5 as the gene of origin with a ClustalW score of 94.

<Alignment>	<Coordinates>	<Matches_Gene>
Pseudogene1_v_JU1424mod	11589-11757	CDS_32
Pseudogene2_v_JU1424mod2	11597-11802	CDS_32

Table 4. Database generated by pseudogenic characterization function of annotation tool.

3.2 Phylogenetic analysis

The MUSCLE and manual alignments of the 84 genomes yielded 3020 unique alignment sequences for RAxML bootstrap analysis. The phylogram generated by Dendroscope from the RAxML maximum-likelihood tree (**Figure 3**) represents evolutionary distances by branch-length.

Large taxa, such as *C. elegans* and *C. briggsae*, were grouped together with high bootstrap values and short evolutionary distances, which is to be expected from members of the same species.

Fig. 3 shows *P. pacificus* segregated from the rest of the *Caenorhabditis* species with a high bootstrap value. As an outgroup, the obvious segregation was expected. *Caenorhabditis sp. 1* also segregated from the rest of the *Caenorhabditis* species. It has been hypothesized that *C. sp. 1* was misclassified into *Ceanorhabditis* genus, which Fig. 3 supports.

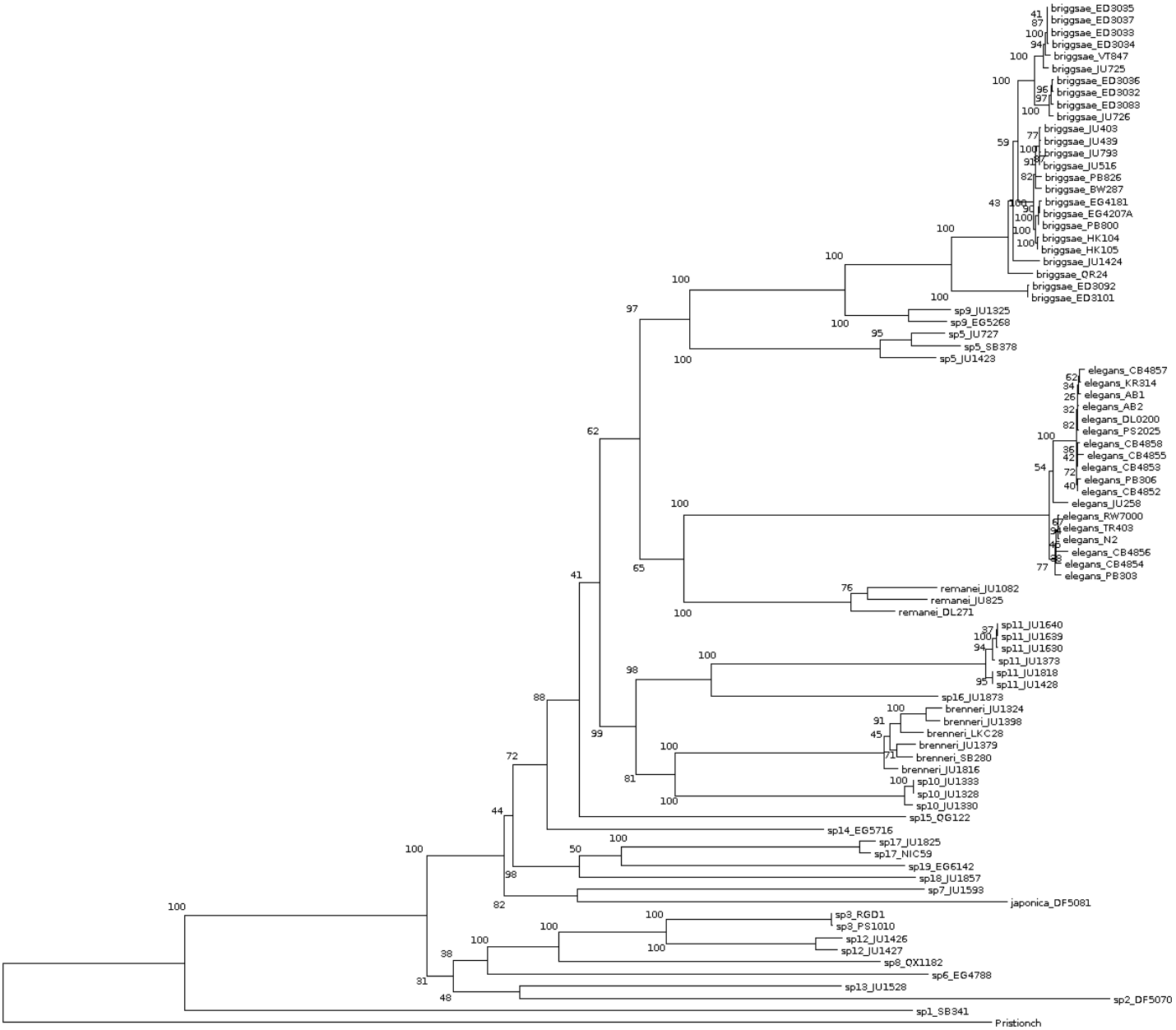

Figure 3. Phylogram of 84 nematode mtDNA genomes across 23 species, including 22 *Caenorhabditis* species and 1 outgroup. The branch-lengths are representative of evolutionary distance. The scores to the left of branches are the bootstrap values. Obtained using Dendroscope (Huson *et al*., 2007).

4 CONCLUSIONS

The automated annotation of *Caenorhabditis* mitochondrial genomes tool has proven its potential to be a powerful resource for not only gene annotation, but also identifying and characterizing intergenic and pseudogenic regions. With more training, the annotation tool will be capable of generating more comprehensive maps of intergenic regions.

The BioPerl module SeqIO proved invaluable for handling FASTA files. NCBI's database handling tools formatdb and FASTAcmd were immensely useful for handling contigs and extracting suspected pseudogenes for alignment.

Careful phylogenetic analyses will need to be conducted to dissect the preliminary findings of the phylogram in Fig. 2. One method would be to conduct the phylogenetic analysis using the amino acid sequences of protein-coding genes. Amino acid sequences are very highly conserved compared to nucleotide sequences due the variability in individual amino acid codons.

6 ACKNOWLEDGEMENTS

I would like to acknowledge the Oregon State University Center for Genome Research and Biocomputing for extensive use of their computing resources. Their computing cluster was used for all of the automated annotation tool development and phylogenetic analysis via RAxML.

7 REFERENCES

C. elegans Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology.** *Science* 1998, 282(5396):2012-2018.

Dorris M, De Ley P, Blaxter ML: **Molecular analysis of nematode diversity and the evolution of parasitism.** *Parasitol Today* 1999, 15(5):188-193.

Gibson T, Blok VC, Dowton M: **Sequence and characterization of six mitochondrial subgenomes from *Globodera rostochiensis*: multipartite structure is conserved among close nematode relatives.** *J Mol Evol* 2007, 65(3):308-315.

Howe DK, Denver DR: **Muller's Ratchet and compensatory mutation in *Caenorhabditis briggsae* mitochondrial genome evolution.** *BMC Evol Bio* 2008, 8:62.

Hu M, Chilton NB, Gasser RB: **The mitochondrial genome of *Strongyloides stercoralis* (Nematoda) - idiosyncratic gene order and evolutionary implications.** *Int J Parasitol* 2003, 33(12):1393-1408.

Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R**. Dendroscope: An interactive viewer for large phylogenetic trees.** *BMC Bioinformatics* 2007 Nov 22, 8(1):460

Hyman BC, Beck JL, Weiss KC: **Sequence amplification and gene rearrangement in parasitic nematode mitochondrial DNA.** *Genetics* 1988, 120(3):707-712.

Kang S, Sultana T, Eom KS, Park YC, Soonthornpong N, Nadler SA, Park JK: **The mitochondrial genome sequence of *Enterobius vermicularis* (Nematoda:**

Oxyurida)--an idiosyncratic gene order and phylogenetic information for chromadorean nematodes. *Gene* 2009, 429(1-2):87-97.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) **ClustalW and ClustalX version 2**. *Bioinformatics* 23, 2947–2948.

Okimoto R., Macfarlane J.L., Clary D.O., Wolstenholme D.R.: **The mitochondrial genomes of two nematodes, *Caenorhabditis elegans* and *Ascaris suum***. *Genetics* 1992, 130:471-498.

Stamatakis A: **RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models**. *Bioinformatics* 2006, 22(21):2688–2690.

Stamatakis A, Hoover P, Rougemont J: **A Rapid Bootstrap Algorithm for the RAxML Web-Servers**. To be published.

Tang S, Hyman BC: **Mitochondrial genome haplotype hypervariation within the isopod parasitic nematode *Thaumamermis cosgrovei***. *Genetics* 2007, 176(2):1139-1150.

5 APPENDIX

5.1 Use of SeqIO module

When the user submits a FASTA file with the desired genomic sequence for annotation via command line, the file is “slurped” into an array. A SeqIO object is created using reference *C. elegans* mitochondrial genes. The “while” loop generates ClustalW input files for each reference *C. elegans* mitochondrial gene versus the user-submitted sequence, then executes ClustalW alignments for each generated file.

```
my $input_file = $ARGV[0]; # User-submitted FASTA file
open(INPUT, $input_file);
my @input = <INPUT>; # Slurps reference file into an array
close INPUT;

my $mt_seq = Bio::SeqIO->new(-file   => "Mt_genes.fa", -format => "FASTA");
# Creates a SeqIO object of mt genes

while (my $seq = $mt_seq->next_seq) {
        my $out_f = $seq->id."_v_".$contigfilename.".fa"; # Creates ClustalW input file
        open(OUT, ">".$out_f); # Opens ClustalW input file
        print OUT @ref; # Prints genomic sequence to ClustalW input file
        print OUT "\n>".$seq->id."\n".$seq->seq."\n"; # Prints mt sequence to input file
        close OUT;
```

```
        my $run_clustal = "clustalw -INFILE=$out_f";
        system($run_clustal); # Runs ClustalW with the newly created input file
    }
```

5.2 Gene coordinate extraction and alignment gap handling

Gene coordinate extraction was accomplished by dismantling the ClustalW-generated alignment file and comparing the genomic sequence to the mitochondrial gene sequence base-by-base. The ClustalW-generated alignment file is “slurped” into an array and broken into lines of the user-submitted sequence and the reference sequence. Each line is broken into individual characters utilizing the “split” function. Aligned sequences are identified when characters in the user-submitted sequence match the characters in the reference mitochondrial sequence.

The code below generalizes the coordinate extraction process:

```
        my @I = split('',$inp_seq); # Splits user-submitted sequence into individual characters
        my @M = split('',$mt_seq); # Splits reference mitochondrial sequence into characters
        my $start, $end, $gap = 0;

for(0..scalar(@I)-1){ # For every character in the genomic sequence
                if ($M[$_] eq '-' && !$start){ # Skip this step if there is no alignment at this point
                        next;
                }
```

```
        if ($M[$_] eq $I[$_] && !$start){ # Defines first coordinate of the gene.
            $start = $_+1;
        }
        if ($M[$_] eq '-' && !$end){ # Defines last coordinate of the gene.
            $end = $_;
        }
```

Hyphens in the mitochondrial sequence denote no alignment. If the automated annotation tool encounters hyphens after an alignment, it defines the final alignment coordinate as the last position it recognized alignment. The automated annotation tool identifies alignment gaps after it encounters an alignment, defines a beginning coordinate, encounters a hyphen, defines an end coordinate and comes across another alignment. The code defines an alignment gap if the end coordinate is defined as a nonzero number and more alignment is identified between the user-submitted and mitochondrial sequences:

```
        if ($M[$_] eq $I[$_] && $end && !$gap){
            $gap_start = ++$end; # Defines the beginning coordinate of the gap
            $gap_end = $_; # Defines ending coordinate of the gap
            $end = 0; # End is null as there is alignment remaining to process
            $gap++;
        }
    }
```