Automated Annotation of *Caenorhabditis* Mitochondrial Genomes and Phylogenetic Analysis

Jessica Campbell[1], Dr. Dee Denver[2]

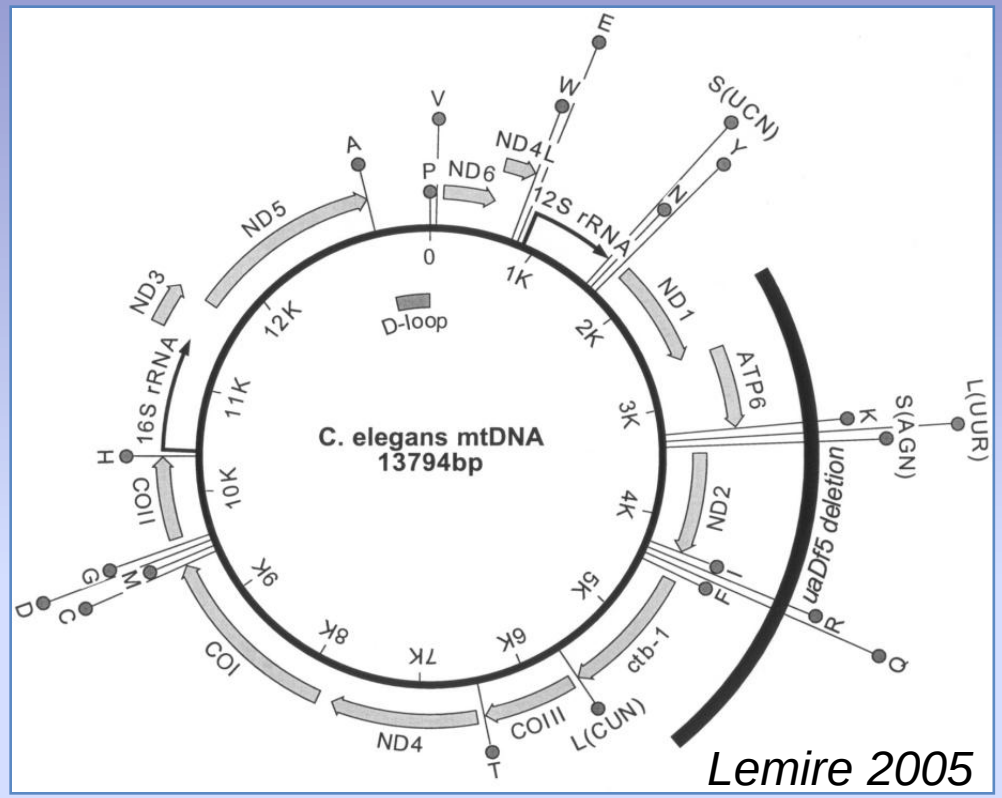
[1]BioResource Research, [2]Department of Zoology

Oregon State University, Corvallis, OR 97331.

Nematodes

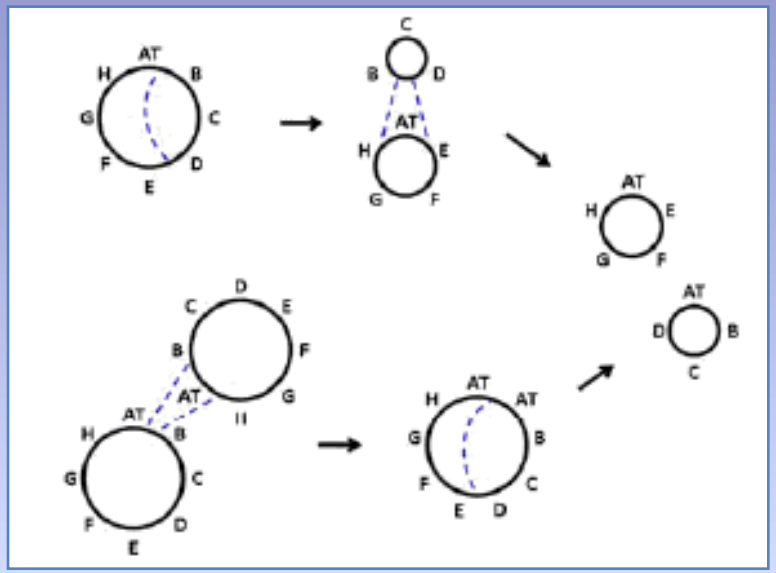
- One of the most diverse animals
- *Caenorhabditis elegans*: important model organism
- Little known regarding the evolutionary mechanisms of mitochondrial genomes

C. elegans Mitochondrial Genome



- 36 genes: 22 transfer RNAs, 12 protein-coding, 2 ribosomal RNAs
- High gene conservation
- High degree of gene order conservation

Nematode Mitogenomic Variation



- Radical gene order rearrangements
- Rapid appearance and disappearance of pseudogenes
- Multi-chromosome mitochondrial DNA
  - Strong evidence in support of inter- and intra-genomic illegitimate recombination

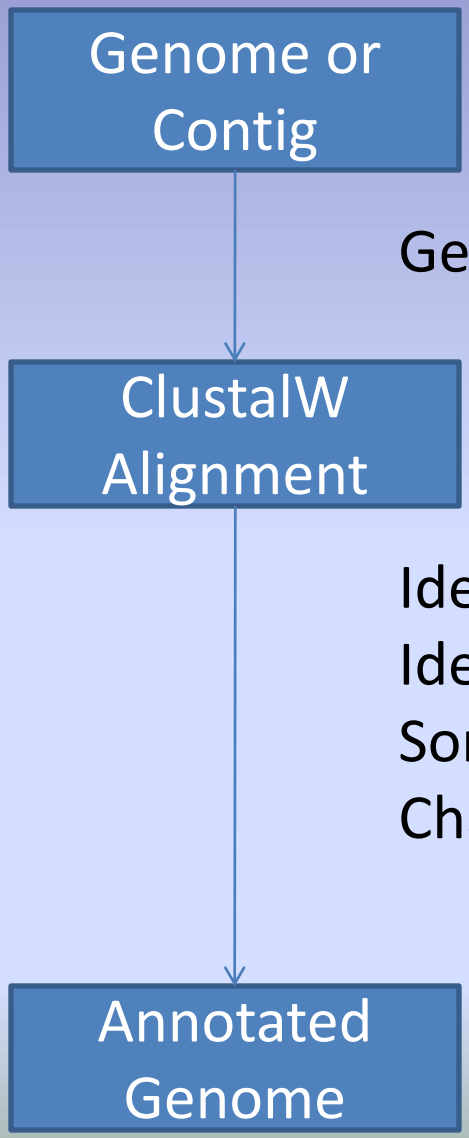
Benefits of Automating Annotation

- Time
- Resource use
- Accessibility

Automated Annotation Developmental Objectives

- Identify gene boundaries
- Classify noncoding regions
- Return:
  - Coordinates for genes in respect to genome
  - Coordinates of noncoding regions
  - Pseudogenic region characterization

Automation Overview



ClustalW Alignment

ClustalW aligns mitochondrial genes against reference genome:

Input: fasta file

```
>DLO200
CAGTAAATAGTTTAATAAAAATATAGCATTTGGGTTGCTAAGATATTATTACTGATAGAATTTTTAGTTTAATTTAGAATGTATCACTTACA$
>tRNA_1
CAGTAAATAGTTTAATAAAAATATAGCATTTGGGTTGCTAAGATATTATTACTGA
```

Output: base-by-base alignment and score

```
CLUSTAL W (1.83) multiple sequence alignment


DLO200          CAGTAAATAGTTTAATAAAAATATAGCATTTGGGTTGCTAAGATATTATTACTGATAGAA
tRNA_1          CAGTAAATAGTTTAATAAAAATATAGCATTTGGGTTGCTAAGATATTATTACTGA-----
                **************************************************** 

DLO200          TTTTTAGTTTAATTTAGAATGTATCACTTACAATGATGGGGTTTAAAATTCTATAGTAAA
tRNA_1          ------------------------------------------------------------
```

```
Sequence format is Pearson
Sequence 1: DLO200         13380 bp
Sequence 2: tRNA_1            55 bp
Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score:  100
```

BioPerl Usage

```
my $mt_seq = Bio::SeqIO->new(-file   => "Mt_genes.fa", -format =>
    "FASTA");
while (my $seq = $mt_seq->next_seq) {
      my $out_f = $seq->id."_v_".$contigfilename.".fa";
      open(OUT, ">".$out_f);
      print OUT @ref;
      print OUT "\n>".$seq->id."\n".$seq->seq."\n“;
      close OUT;

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

Positive Control: *C. elegans* DL0200

<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	

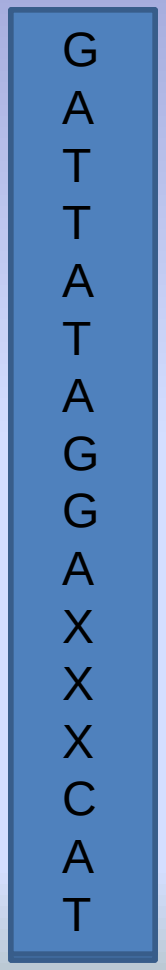
Negative Control: *C. briggsae* JU1424

<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	

Pseudogene Characterization

- Excise sequences larger than 200nt between genes
- Align excised sequence against genomic sequence using ClustalW
- Match alignment coordinates to previously generated gene coordinates

Pseudogene Classification Algorithm



1. Excise intergenic sequence

2. Realign intergenic sequence on modified genomic sequence

3. Determine gene of origin via coordinates

C. briggsae Pseudogenes

<Alignment>	<Coordinates>	<Score>	<Matches_Gene>
Pseudogene1_v_JU1424mod1.aln	11589-11757	96	CDS_32
Pseudogene2_v_JU1424mod2.aln	11597-11802	94	CDS_32

CDS_32_v_JU1424.aln	11573-11908

C. briggsae Pseudogenes

- Originate from nad5, NADH dehydrogenase 5 (CDS_32)
  - ψnad5-1 located between tRNA 6 and tRNA 7
  - ψnad5-2 located between tRNA 31 and nad5 (CDS_32)

Howe and Denver 2008

Future Developments

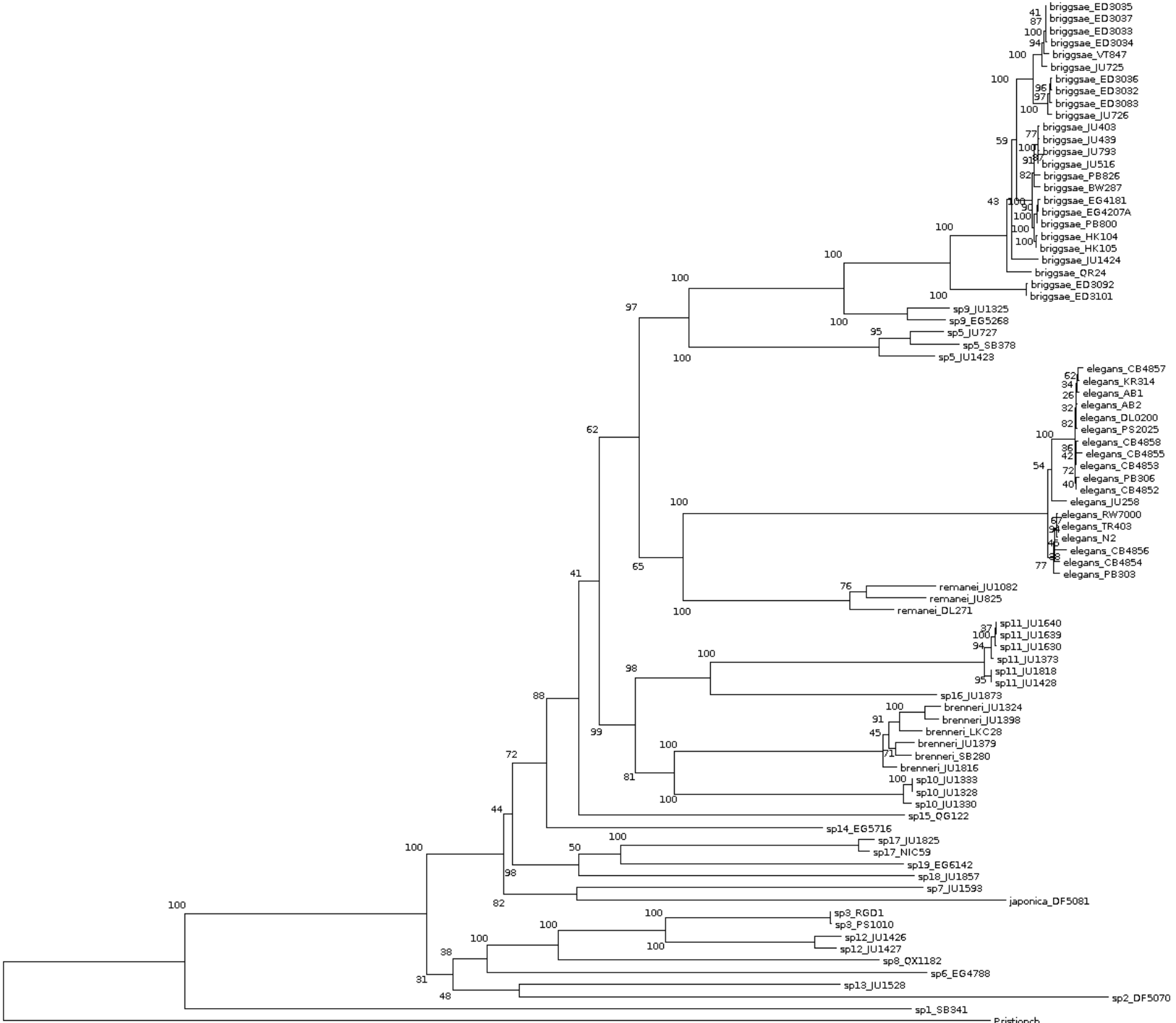
- Phylogenetic analysis of amino acid sequences
- Public access
  - User-friendly front-end
  - Increased internal robustness

Phylogenetic Analysis

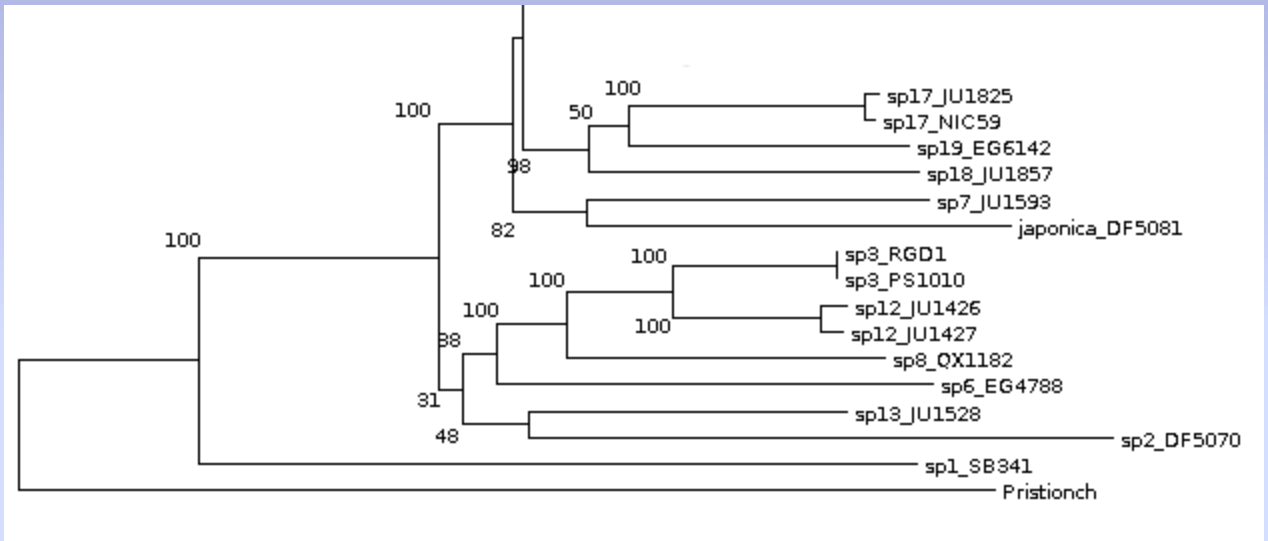
- Explore evolutionary relationships
- 84 sequences: 83 *Caenorhabditis*, 1 outgroup
  - Aligned using MUSCLE
  - Alignment gaps greater than 5nt manually removed in MEGA
  - File converted using ReadSeq
  - Bootstrap analysis and maximum-likelihood tree using RAxML (General Time Reversible model of nucleotide substitution, I' model of heterogeneity; 1000 replicates)
  - Graphical phylogram rendered by Dendroscope

Mitochondrial Sequences Analyzed

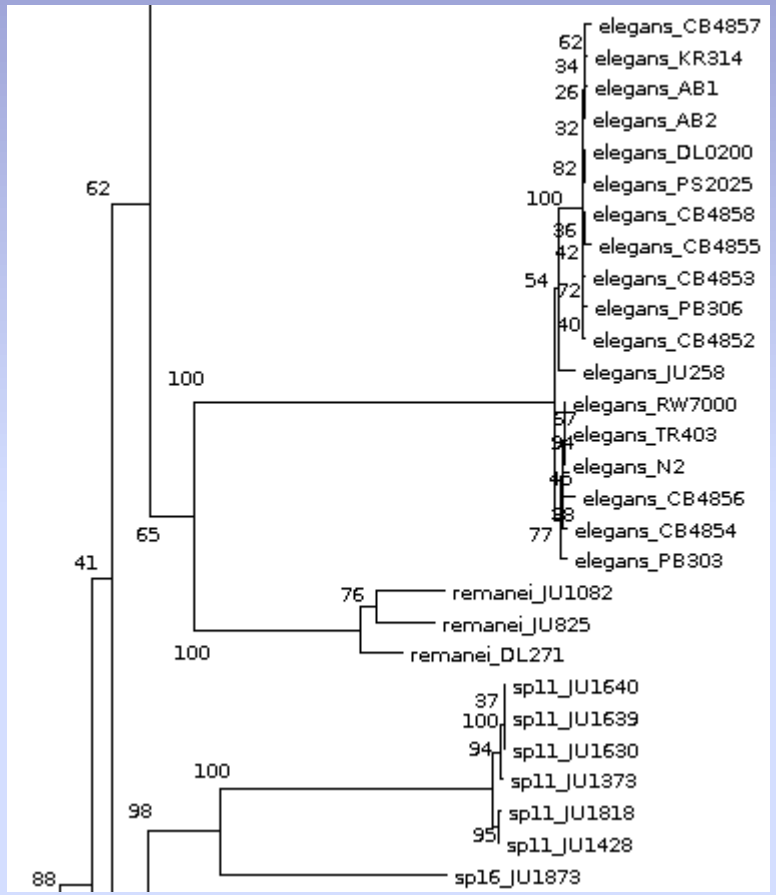
Caenorhabditis briggsae (BW287, ED3032, ED3033, ED3034, ED3035, ED3036, ED3037, ED3083, ED3092, ED3101, EG4181, EG4207A, HK104, HK105, JU403, JU439, JU516, JU725, JU726, JU793, JU1424, OR24, PB800, PB826 and VT847), *Caenorhabditis elegans* (AB1, AB2, CB4852, CB4853, CB4854, CB4855, CB4856, CB4857, CB4858, DL0200, JU258, KR314, N2, PB303, PB306, PS2025, RW7000 and TR403), *Caenorhabditis brenneri* (JU1324, JU1379, JU1398, JU1816, LKC28 and SB280), *Caenorhabditis remanei* (DL271, JU825, and JU1082), *Caenorhabditis japonica* (DF5081), *Caenorhabditis sp. 1* (SB341), *Caenorhabditis sp.* 3 (RGD1 and PS1010), *Caenorhabditis sp. 5* (JU737, JU1423 and SB375), *Caenorhabditis sp. 6* (EG4788), *Caenorhabditis sp. 7* (JU1593), *Caenorhabditis sp. 8* (QX1182), *Caenorhabditis sp. 9* (EG5268 and JU1325), *Caenorhabditis sp. 10* (JU1328, JU1330 and JU1333), *Caenorhabditis sp. 11* (JU1373, JU1428, JU1630, JU1639, JU1640 and JU1818), *Caenorhabditis sp. 12* (JU1426 and JU1427), *Caenorhabditis sp. 13* (JU1528), *Caenorhabditis sp. 14* (EG5716), *Caenorhabditis sp. 15* (OG122), *Caenorhabditis sp. 16* (JU1873), *Caenorhabditis sp. 17* (JU1825 and NIC59), *Caenorhabditis sp. 18* (JU1857), *Caenorhabditis sp. 19* (EG6142) and *Pristionchus pacificus* was used as an outgroup.


briggsae_ED3035
briggsae_ED3037
briggsae_ED3033
briggsae_ED3034
briggsae_VT847
briggsae_JU725
briggsae_ED3036
briggsae_ED3082
briggsae_ED3083
briggsae_JU726
briggsae_JU403
briggsae_JU439
briggsae_JU793
briggsae_JU516
briggsae_PB826
briggsae_BW287
briggsae_EG4181
briggsae_EG4207A
briggsae_PB800
briggsae_HK104
briggsae_HK105
briggsae_JU1424
briggsae_QR24
briggsae_ED3092
briggsae_ED3101
sp9_JU1325
sp9_EG5268
sp5_JU727
sp5_SB378
sp5_JU1423
elegans_CB4857
elegans_KR314
elegans_AB1
elegans_AB2
elegans_DL0200
elegans_PS2025
elegans_CB4858
elegans_CB4855
elegans_CB4853
elegans_PB306
elegans_CB4852
elegans_JU258
elegans_RW7000
elegans_TR403
elegans_N2
elegans_CB4856
elegans_CB4854
elegans_PB303
remanei_JU1082
remanei_JU825
remanei_DL271
sp11_JU1640
sp11_JU1639
sp11_JU1630
sp11_JU1373
sp11_JU1818
sp11_JU1428
sp16_JU1873
brenneri_JU1324
brenneri_JU1398
brenneri_LKC28
brenneri_JU1379
brenneri_SB280
brenneri_JU1816
sp10_JU1333
sp10_JU1328
sp10_JU1330
sp15_QG122
sp14_EG5716
sp17_JU1825
sp17_NIC59
sp19_EG6142
sp18_JU1857
sp7_JU1593
japonica_DF5081
sp3_RGD1
sp3_PS1010
sp12_JU1426
sp12_JU1427
sp8_QX1182
sp6_EG4788
sp13_JU1528
sp2_DF5070
sp1_SB341

Phylogram Results



- Species grouped appropriately
- Outgroup segregates as anticipated


elegans_CB4857
elegans_KR314
elegans_AB1
elegans_AB2
elegans_DL0200
elegans_PS2025
elegans_CB4858
elegans_CB4855
elegans_CB4853
elegans_PB306
elegans_CB4852
elegans_JU258
elegans_RW7000
elegans_TR403
elegans_N2
elegans_CB4856
elegans_CB4854
elegans_PB303
remanei_JU1082
remanei_JU825
remanei_DL271
sp11_JU1640
sp11_JU1639
sp11_JU1630
sp11_JU1373
sp11_JU1818
sp11_JU1428
sp16_JU1873
62
34
26
32
82
100
36
42
54
72
40
100
65
41
76
100
37
100
94
100
98
95
88

Future Research

- Phylogenetic analyses using amino acid sequences
- Use of GTRCAT experimental treebuilding substitution matrices in RAxML

Acknowledgements

- Larry Wilhelm
- Cloud computing resources at Oregon State University's Center for Genome Research and Biocomputing

References