Nematode gene annotation by machine learning assisted proteotranscriptomics enables proteome-wide evolutionary analysis

Supplemental Data

Supplemental Figures S1-S12

Supplemental Tables S1-S10

Supplemental Material (.zip folder)

Protein_Ortho_table.zip: Table of all orthology groups established by ProteinOrtho based on proteotranscriptomics annotations of all 12 species.

Transdecoder_ORF_prediction_cds_seq_raw.zip: CDS sequence FASTA files of all TransDecoder ORF predictions for all 12 species (separate files for genome-guided and genome-free transcriptome assembly).

Transdecoder_ORF_predictions_cds_seq_evidenced.zip: CDS sequence FASTA files of TransDecoder ORF predictions with mass spectrometry peptide evidence for all 12 species (separate files for genome-guided and genome-free transcriptome assembly).

Transdecoder_ORF_predictions_pep_seq_raw.zip: protein sequence FASTA files of TransDecoder ORF predictions with mass spectrometry peptide evidence for all 12 species (separate files for genome-guided and genome-free transcriptome assembly).

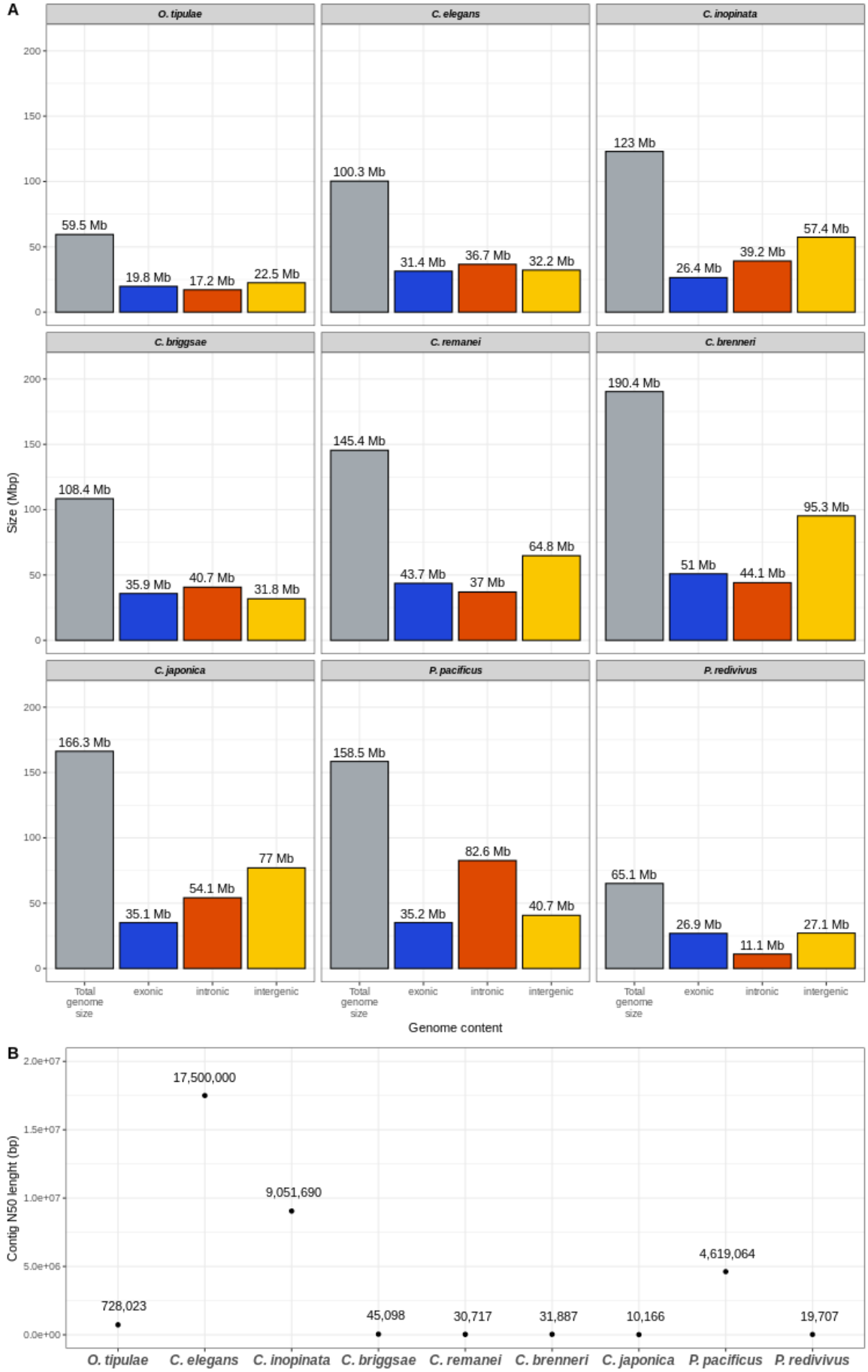
Transdecoder_ORF_predictions_pep_seq_evidenced.zip: protein sequence FASTA files of TransDecoder ORF predictions with mass spectrometry peptide evidence for all 12 species (separate files for genome-guided and genome-free transcriptome assembly).

Trinity_assemblies.zip: Transcript sequence FASTA files of Trinity assembled transcripts for all 12 species (separate files for genome-guided and genome-free transcriptome assembly).

Trinotate.zip: Tables of Trinotate annotations of all Trinity assembled transcripts for all 12 species (separate files for genome-guided and genome-free transcriptome assembly).

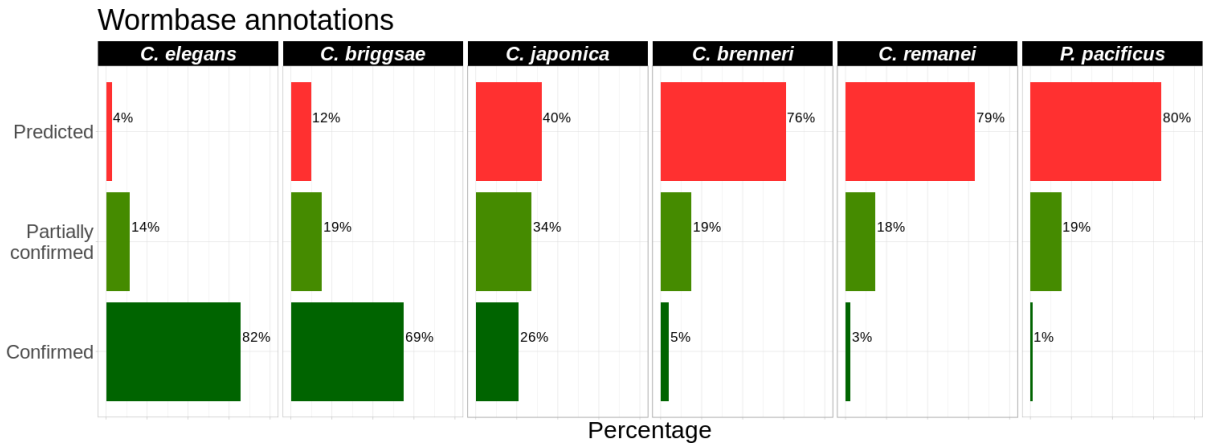
Blast_results_new_genes.pdf: BLASTP results for the two new *C. elegans* ORFs.

Supplemental Figure S1



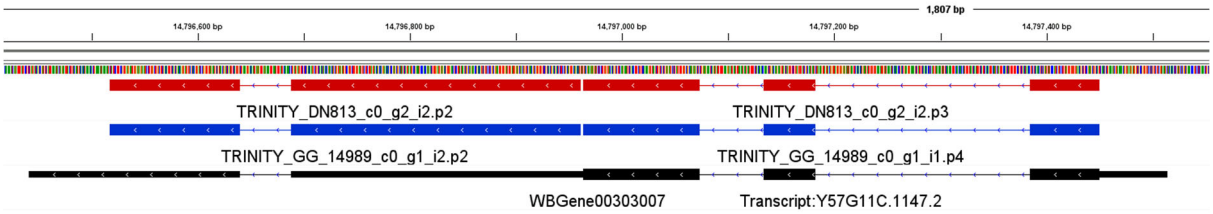
Supplemental Figure S1. Genome content and genome assembly contig N50 length of 9 nematodes. (A) Genome content extracted from WormBase genome assembly and gene annotation files (version WS273). Bar plots show the total genome size in gray and the proportions of exonic (blue), intronic (orange), and intergenic (yellow) regions for all nine species that have genome assemblies available. As the data was extracted from assemblies of varying quality (see Supplemental Table S1) there is no warranty of the accuracy of these distributions. (B) Genome contiguity of all species that have genome assemblies available in WormBase (version WS273) plotted as contig N50 lengths.

Supplemental Figure S2



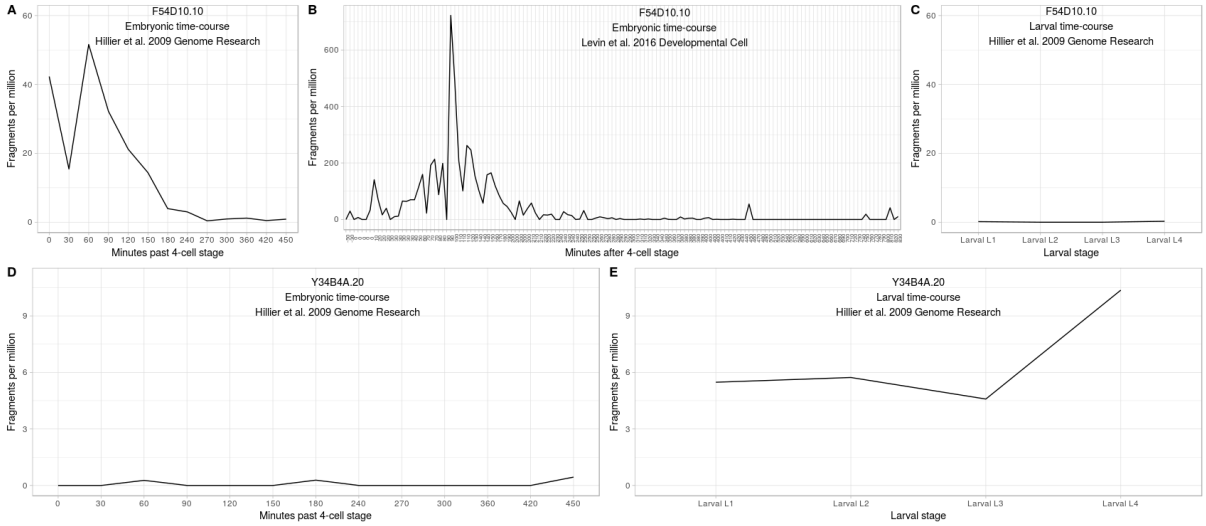
Supplemental Figure S2. Distribution of WormBase annotation confirmation level across all *Caenorhabditis* species. Categories are (1) predicted (red) - unsupported gene predictions, (2) partially confirmed (light green) - not all parts of the ORF are confirmed, and (3) confirmed (dark green) - all parts, translation start and stop site, all coding exons, and exon/intron junctions are confirmed by experimental data.

Supplemental Figure S3



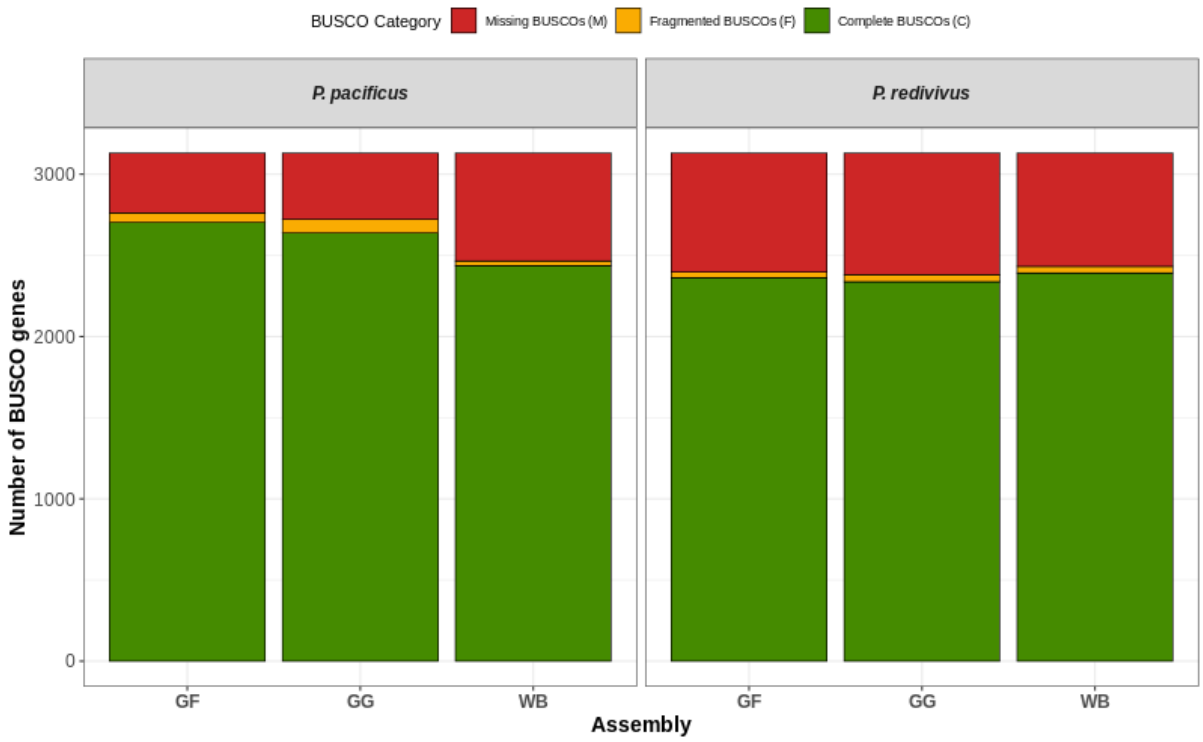
Supplemental Figure S3. Visualization of an example of a fragmented gene model via Integrative Genomics Viewer (IGV) browser. *C. elegans* GF (red) and GG (blue) assembled transcripts were mapped to the *C. elegans* genomic sequence and are shown side by side with the respective *C. elegans* WormBase entry (black) on Chromosome IV.

Supplemental Figure S4



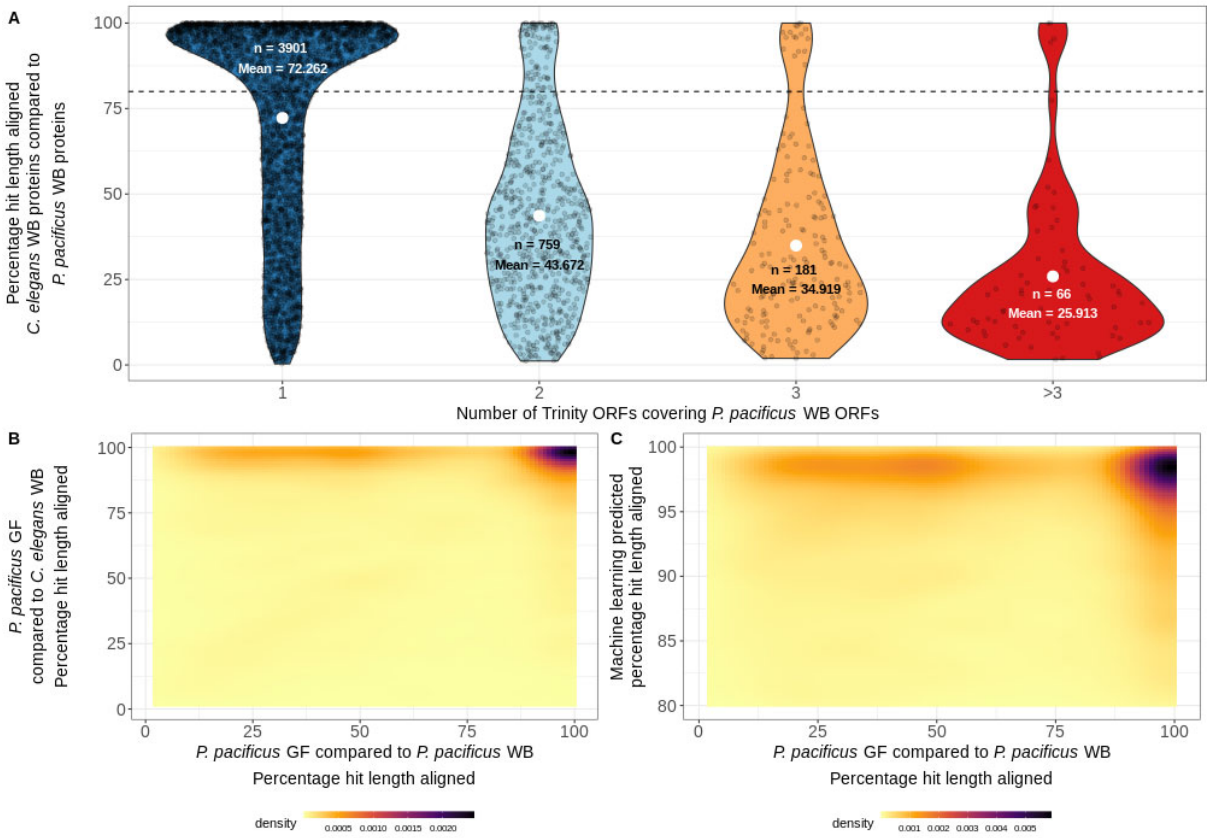
Supplemental Figure S4. Expression profiles of F54D10.10 and Y34B4A.20 during developmental stages of *C. elegans*. (A) F54D10.10 transcript shows expression during embryonic developmental time-course. (B) Validation of F54D10.10 embryonic expression in an additional embryonic transcriptome time-course. (C) F54D10.10 expression at the 4 larval stages. (D) Y34B4A.20 has no expression during early embryonic development. (E) Y34B4A.20 shows increased expression at the L4 stage.

Supplemental Figure S5



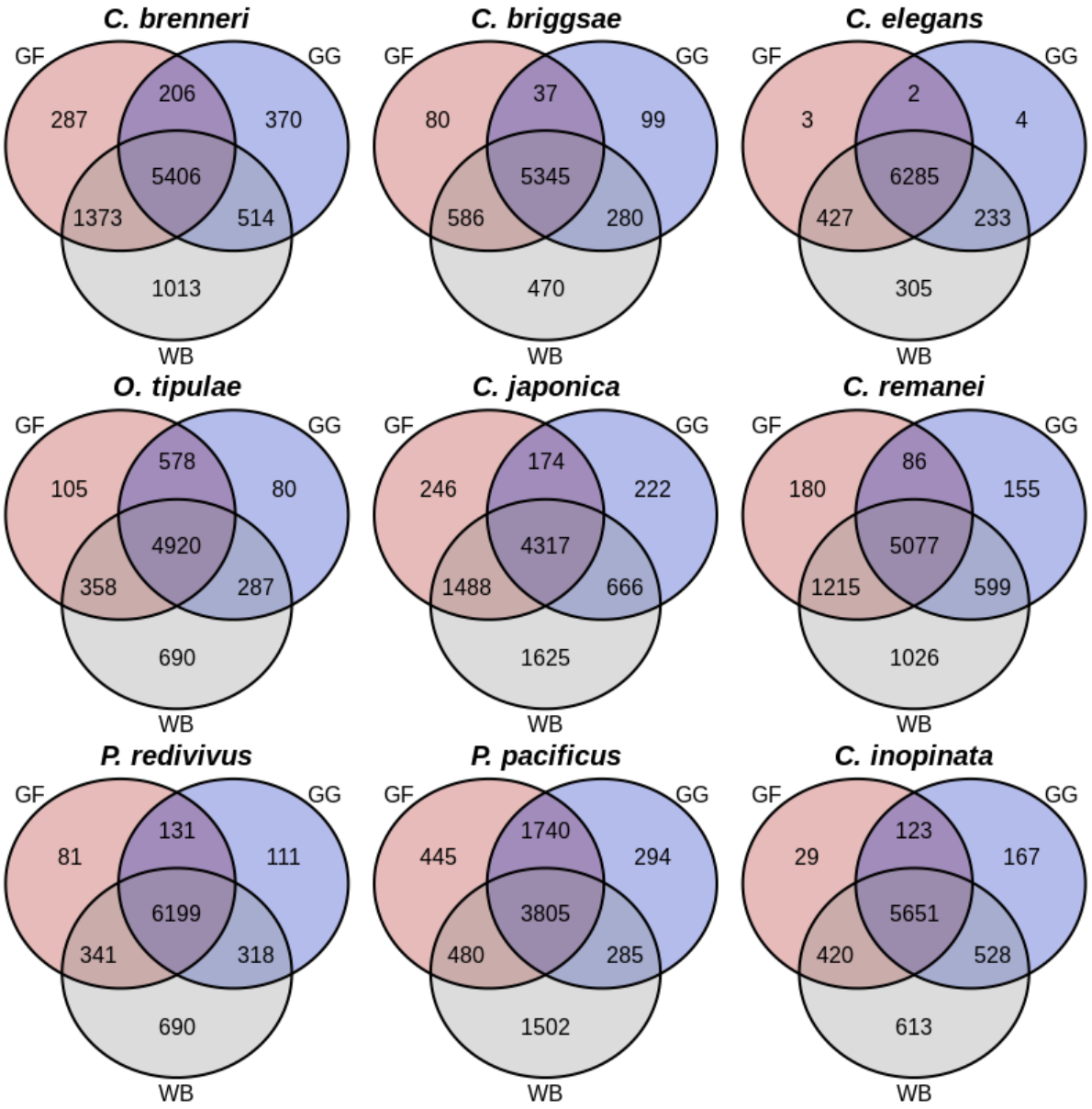
Supplemental Figure S5. Results of BUSCO analysis comparing GF and GG assemblies with the current WormBase annotation of *P. pacificus* and *P. redivivus*. The y-axis represents the counted number of BUSCO genes and the x-axis shows different evaluated assemblies. Green: complete and single-copy genes; orange: fragmented genes; red: missing genes, showing that the absence is not an artifact of our methodology.

Supplemental Figure S6



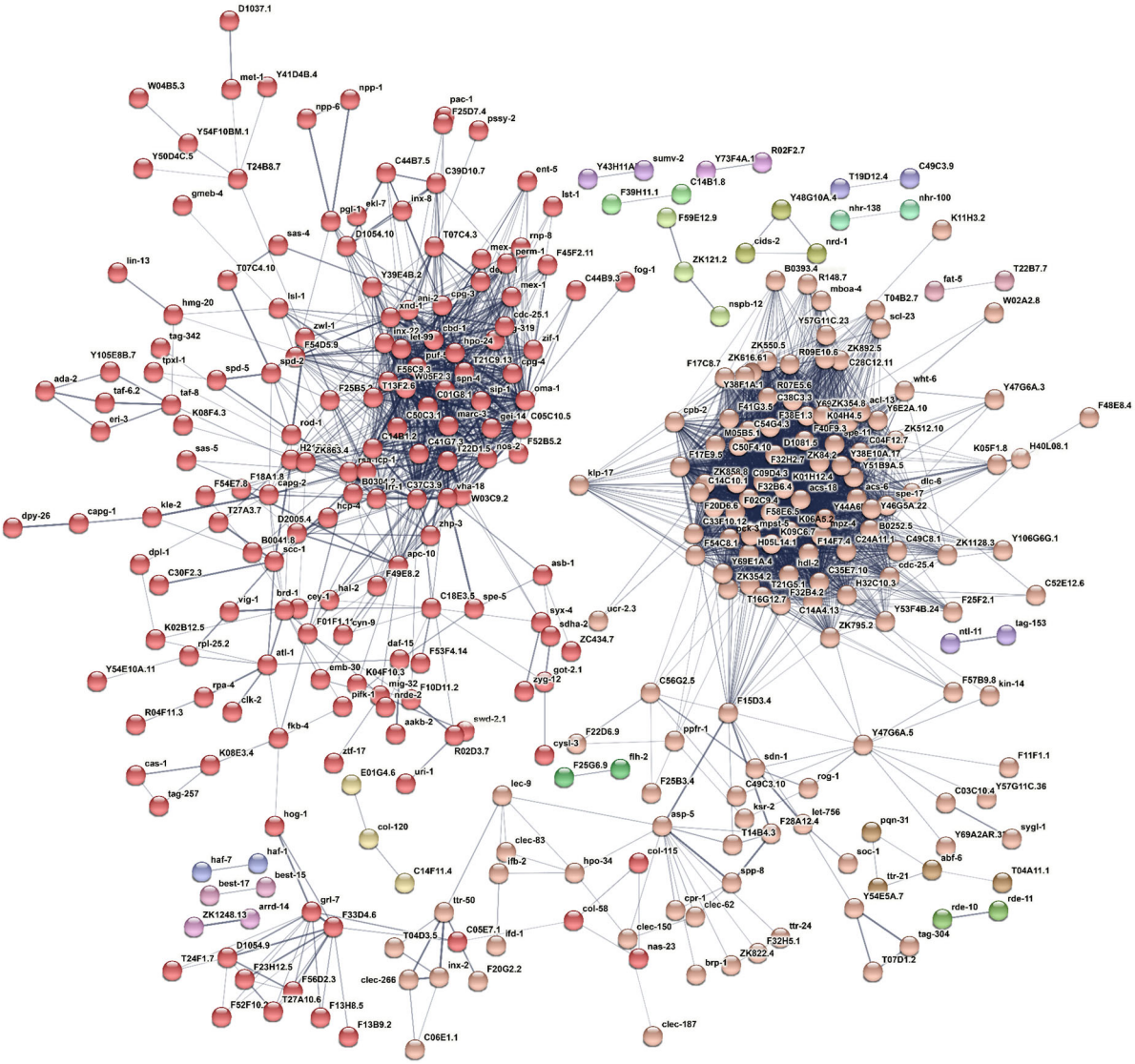
Supplemental Figure S6. Validation of *P. pacificus* fusion bias. (A) Percentage hit length of WormBase *P. pacificus* proteins when compared to WormBase *C. elegans* orthologs established by blastp. *P. pacificus* proteins were grouped by the number of proteins needed to cover the same protein sequence in the Trinity genome-free (GF) assembly (1, 2, 3 or more proteins, same protein sets and color code as in main Figure 2a). While Wormbase annotated *P. pacificus* proteins that are coherent with the GF annotated proteins (overlap with only one GF annotated protein) show high percentage hit lengths with WormBase *C. elegans* proteins, this value decreases significantly for proteins that have signals of falsely predicted fusion (WormBase proteins with more than one overlapping protein from GF). (B) 2-D kernel density plot of the percentage hit length of *P. pacificus* GF proteins when compared to *P. pacificus* WormBase and to the *C. elegans* WormBase annotation. The cloud in the upper left corner clearly shows GF assembled proteins that seem to be fragmented when compared to the WormBase *P. pacificus* proteins; however, these proteins show high percentage hit length when compared to *C. elegans* WormBase annotations and hence probably represent artifacts in the current *P. pacificus* annotation. (C) 2-D kernel density plot of the percentage hit length of *P. pacificus* GF proteins when compared to *P. pacificus* WormBase and the predicted percentage hit length based on our machine learning algorithm. The cloud in the upper left corner again clearly shows GF assembled proteins that seem to be fragmented when compared to the WormBase *P. pacificus* proteins, however, show high machine learning established percentage hit length and might indeed represent artifacts in the *P. pacificus* WormBase annotation.

Supplemental Figure S7



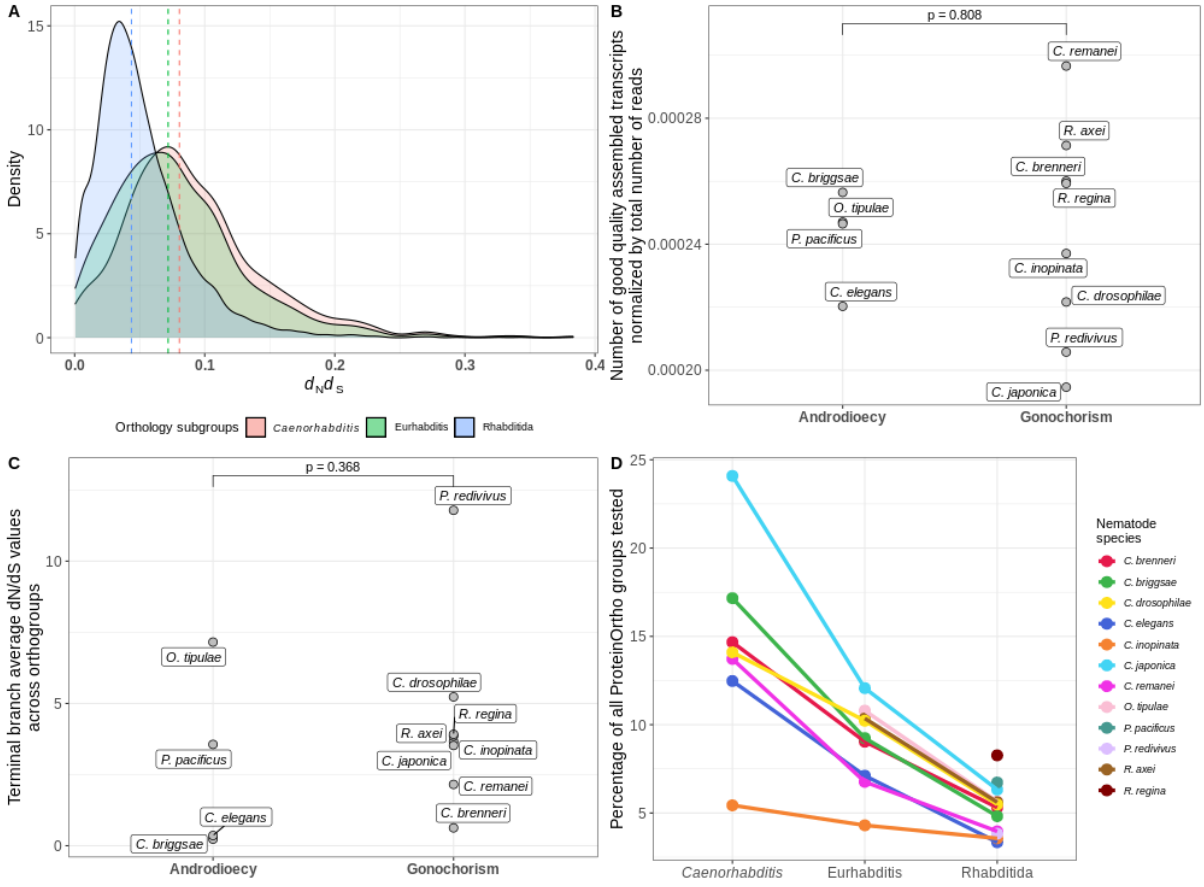
Supplemental Figure S7. Venn diagrams depicting the overlap between the identified proteins of WB (WormBase in gray), GF (genome-free in red), and GG (genome-guided in blue) proteomes for each studied species.

Supplemental Figure S8



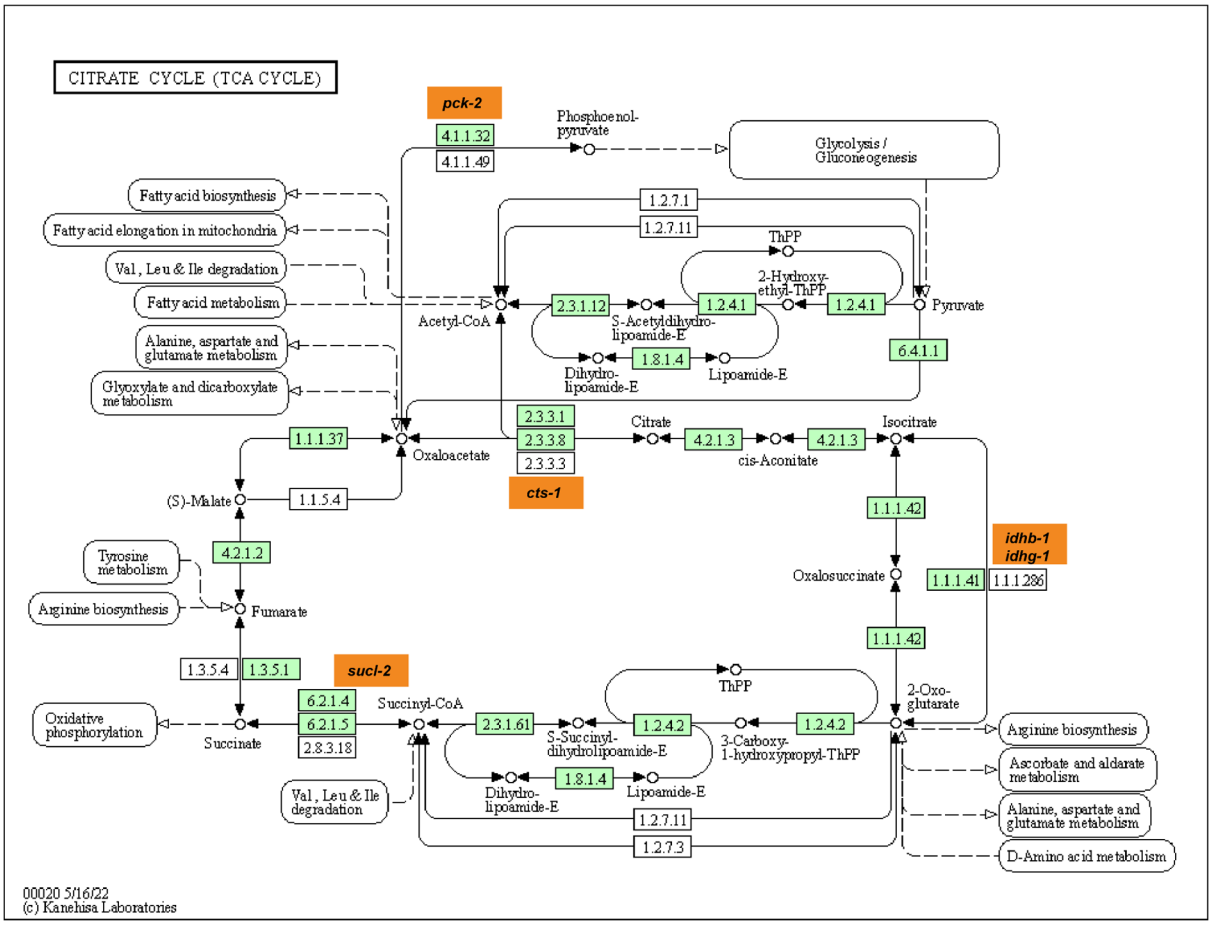
Supplemental Figure S8. STRINGdb network plot of detected genes specific to the genus *Caenorhabditis*. Nodes represent *C. elegans* proteins with orthologs only in *Caenorhabditis* species (absent from all other studied species) and edges represent protein-protein associations provided by STRINGdb. Node colors distinguish association clusters based on MCL clustering (see also Supplemental Table 4).

Supplemental Figure S9



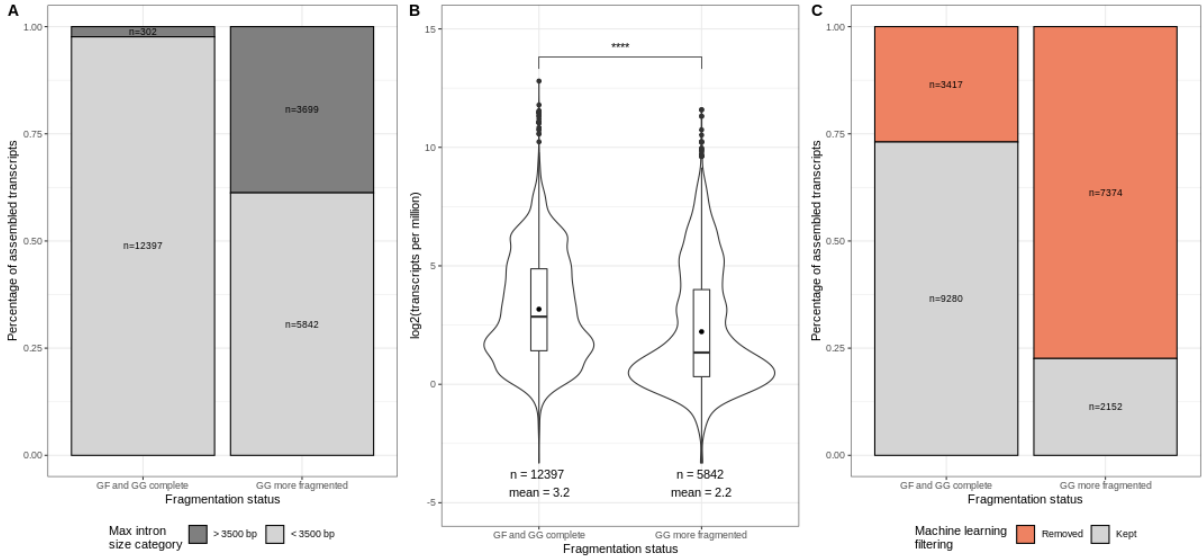
Supplemental Figure S9. Different levels of adaptive evolution detected in a phylogeny of nematodes. (A) Density plot of M0 model d_N/d_S (ω) values calculated for 5,417 orthologs in 12 nematodes species, evaluated for Rhabditida (blue), Eurhabditis (green), and *Caenorhabditis* (red). The median of each group is represented with a dashed line. All distributions show high levels of purifying selection (ω>0) in the majority of the codon sites. The differences in the medians and shift in the distributions of the values between the different groups emphasize the decrease in the detection sensitivity of adaptive evolution with an increasing degree of divergence between species (*Caenorhabditis* > Eurhabditis > Rhabditida). (B) Assembly efficiency measured as the number of assembled transcripts that pass the machine learning completeness prediction of 80% normalized by the total number of sequenced reads used for the assembly is shown for species divided into gonochoristic and androdioecious mode of reproduction. Due to missing genome annotations and uncertainty regarding the quality of some of the existing assemblies only genome-free assembled transcripts are represented. (C) Terminal branch average d_N/d_S values across 1-to-1 orthogroups are shown for species divided into gonochoristic and androdioecious mode of reproduction. (D) Percentages of orthologous groups under positive selection, grouped by subsets of species included in the analysis - Rhabditida, Eurabditis, and *Caenorhabditis*.

Supplemental Figure S10



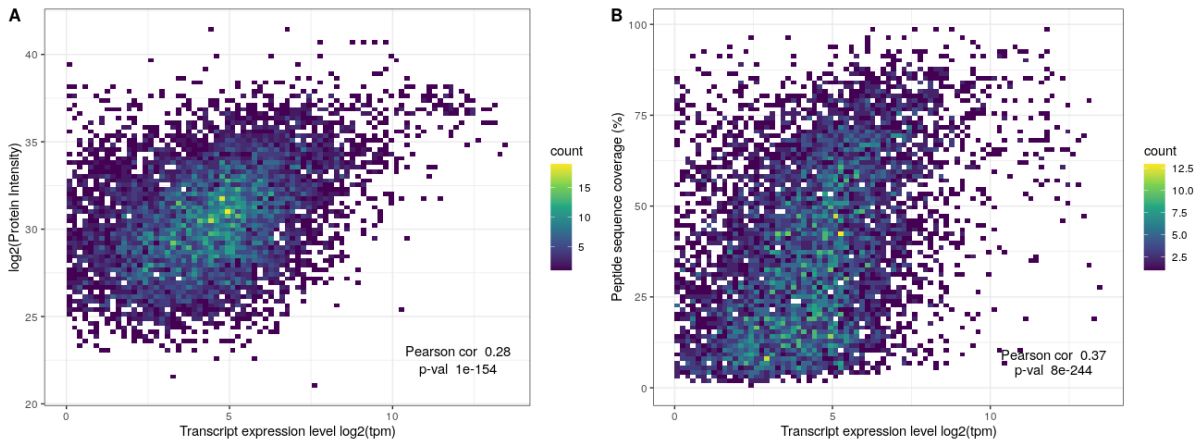
Supplemental Figure S10. Tricarboxylic acid cycle (TCA) KEGG pathway. TCA cycle genes under positive selection in *C. japonica* are highlighted in orange. Pathway diagram was adapted from https://www.kegg.jp/pathway/cel00020.

Supplemental Figure S11



Supplemental Figure S11. Analyses of genome-guided dependent biases in *C. elegans*. (A) Stacked barplot showing the proportions of *C. elegans* transcripts that contain introns shorter (light gray) or longer (dark gray) than 3500 bases in the group of transcripts that are complete in the genome-free (GF) and the genome-guided (GG) assembly in comparison to those that show fragmentation in GG. (B) Violin plots showing the distribution of the expression levels of *C. elegans* transcripts in the group of transcripts that are complete in the genome-free (GF) and the genome-guided (GG) assembly in comparison to those that show fragmentation in GG. (C) Stacked barplot showing the proportions of *C. elegans* transcripts that were either filtered out or passed the threshold of 80% completeness as predicted by the applied machine learning completeness prediction in the groups of transcripts that are complete in the genome-free (GF) and the genome-guided (GG) assembly in comparison to those that show fragmentation in GG.

Supplemental Figure S12



Supplemental Figure S12. Correlation between transcript level and protein intensity and peptide sequence coverage. (A) Density plot of protein intensities measured as a function of the respective transcript expression level at the transcriptome level measured by RNA-seq. (B) Density plot of peptide sequence coverage percentage detected as a function of the respective transcript expression level at the transcriptome level measured by RNA-seq.