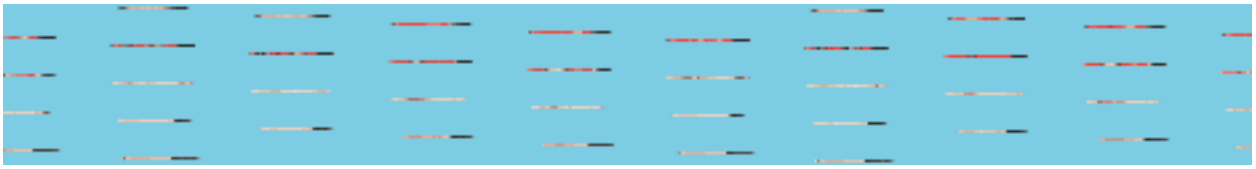
Nematode gene annotation by machine-learning-assisted proteotranscriptomics enables proteome-wide evolutionary analysis

Alejandro Ceron-Noriega, Miguel V. Almeida, Michal Levin, et al.

Genome Res. 2023 33: 112-128 originally published online January 18, 2023
Access the most recent version at doi:10.1101/gr.277070.122



Method

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

Alejandro Ceron-Noriega,[1] Miguel V. Almeida,[1,3] Michal Levin,[1,2] and Falk Butter[1,2]

[1]*Institute of Molecular Biology (IMB), 55128 Mainz, Germany*

Nematodes encompass more than 24,000 described species, which were discovered in almost every ecological habitat, and make up >80% of metazoan taxonomic diversity in soils. The last common ancestor of nematodes is believed to date back to ~650–750 million years, generating a large and phylogenetically diverse group to be explored. However, for most species high-quality gene annotations are incomprehensive or missing. Combining short-read RNA sequencing with mass spectrometry–based proteomics and machine-learning quality control in an approach called proteotranscriptomics, we improve gene annotations for nine genome-sequenced nematode species and provide new gene annotations for three additional species without genome assemblies. Emphasizing the sensitivity of our methodology, we provide evidence for two hitherto undescribed genes in the model organism *Caenorhabditis elegans*. Extensive phylogenetic systems analysis using this comprehensive proteome annotation provides new insights into evolutionary processes of this metazoan group.

[Supplemental material is available for this article.]

Nematodes are one of the most diverse, abundant, and widespread metazoan phylum on earth (Bongers and Bongers 1998; Hodda et al. 2009; Blaxter 2016). They inhabit a broad range of ecological niches with lifestyles ranging from free-living to plant and animal parasitic, including varying reproduction modes, morphology, and developmental programs (Kiontke and Fitch 2013; Vlaar et al. 2021). Nematodes account for over three-quarters of all individual animals on the planet, encompassing 24,000 described and 1 million estimated existing species (Blaxter 2016), including the important model organism *Caenorhabditis elegans*, which has been introduced to the laboratory in the early 1970s (Brenner 1973). *C. elegans* has been extensively studied for almost half a century as a model for development, neurobiology, disease progression, and aging (Horvitz 2003; Kaletta and Hengartner 2006; Antoshechkin and Sternberg 2007; Leung et al. 2008). Because of its importance, it was the first fully assembled animal genome with a comprehensive, well-evidenced, and high-quality gene annotation. Two other species of its genus also have well-annotated genomes and are especially used for evolutionary comparisons: (1) *Caenorhabditis briggsae* (Hillier et al. 2005), the satellite species of *C. elegans*, which shares remarkable similarity in morphology and developmental programs (Gupta et al. 2007), being genomically as divergent from *C. elegans* as human from mouse (Cutter 2008); and (2) the recently identified sister species of *C. elegans* termed *Caenorhabditis inopinata* (Kanzaki et al. 2018). Comparisons between genomes of different species can provide insights into genetic pathways, which in combination with ecological information deliver clues to how organisms adapt to their environment (Stevens et al. 2019). To enable broader evolutionary comparisons, a larger set of well-annotated species would be highly beneficial. For a more comprehensive picture of the genome evolution among nematodes, the community has provided additional genome assemblies, for example, for *Caenorhabditis brenneri, Caenorhabditis japonica, Caenorhabditis remanei,* and *Pristionchus pacificus*, accessible in databases like WormBase (Howe et al. 2012; Harris et al. 2020). These genome assemblies encompass a wide variety of genome sizes and compactness (Supplemental Fig. S1A). However, the quality of these assemblies is not uniform, and some show highly fragmented contigs and gaps (Supplemental Table S1; Supplemental Fig. S1B), rendering global estimations vague. Unfortunately, assembly quality plays a major role in the accuracy of ab initio gene prediction; that is, mistakes in genome assemblies can lead to the erroneous addition and/or subtraction of gene annotations (Han et al. 2013). Thus far, as most of the nematode annotations are still based on automated annotation pipelines and not on experimental evidence (Supplemental Fig. S2), the gene prediction quality cannot be estimated confidently. This represents a bottleneck in the broad-scale understanding of nematode evolution and may lead to misinterpretations (Han et al. 2013). As a result, evolutionary studies across different species have so far focused primarily on the detection of selection signatures at the single-gene family level (Thomas et al. 2005; Thomas 2006; Mukherjee and Bürglin 2007; Weinstein et al. 2019). To enable accurate orthology assignment and allow for extensive investigations of the evolution in this phylum, experimentally validated annotations are essential. To address this issue, de novo assembled contigs from RNA-seq data of poly(A)-enriched mRNA are useful. As mRNAs are devoid of introns, the resulting predictions are likely more accurate than predicting gene models from genomic sequences that are based on ambiguous splice-site predictions. Furthermore, protein-coding gene validation by

[2]These authors contributed equally to this work.
[3]Present address: Wellcome Trust/Cancer Research UK Gurdon Institute and Department of Genetics, University of Cambridge, Cambridge CB2 1QN, UK
Corresponding authors: m.levin@imb.de, f.butter@imb.de
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.277070.122.

33:112–128 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/23; www.genome.org

additional peptide evidence through high-resolution mass spectrometry can strongly improve the annotation as shown in various previous studies investigating individual species (Jaffe et al. 2004; Castellana et al. 2008; Desgagné-Penix et al. 2010; Evans et al. 2012; Volkening et al. 2012; Mohien et al. 2013; Kumar et al. 2016; Chapman and Bellgard 2017; Prasad et al. 2017; Ma et al. 2018; Lang et al. 2019; Ding et al. 2020; Levin et al. 2020; Müller et al. 2021).

Here we use an automated, systematic, generic, and scalable proteotranscriptomics assembly (PTA) workflow (Levin and Butter 2022) for high-confidence annotation of protein-coding genes in 12 nematode species. Including a novel machine-learning implementation to score transcript fragmentation, which is a well-known issue of transcriptome assemblies (Treangen and Salzberg 2011), we improve existing annotations for nine genome-sequenced nematode species and provide annotations for three species that currently lack genome assemblies enabling broad evolutionary analyses.

Results

Benchmark of de novo transcriptome assembly

As we aimed to provide and interrogate extensive protein-coding gene annotations for a broad range of nematodes, including species with low-quality or nonexistent genome assemblies, we decided on a de novo approach using assembled contigs from RNA-seq data of poly(A)-enriched mRNA. We selected *C. elegans*, the best-annotated nematode species, for benchmarking the quality and completeness of our chosen transcriptome assembly approach. We thus generated 74 million paired-end RNA-seq reads of 79-base length from a nonsynchronized *C. elegans* culture containing all developmental stages ranging from embryos to adult worms. The RNA-seq reads were quality controlled and either used for genome-free (GF) or genome-guided (GG) transcriptome assembly with the Trinity suite (Grabherr et al. 2011). For the GF approach, reads are directly assembled, whereas in the GG approach, reads are first mapped to the genome and then assembled into contigs considering mapping information. TransDecoder (Haas et al. 2013) predicted 39,538 potential open reading frames (ORFs) for the GF assembly and 41,509 ORFs for the GG assembly. The 50th percentile lengths (N50) of transcripts with very high expression levels (Ex90N50) were 2467 nt for GF and 2343 nt for GG. It is noteworthy that the N50 values of each expression bin (ExN50) were highly similar to the ExN50 values of the *C. elegans* WormBase annotation, especially for the GF assembly (Pearson's r of 0.96 for GF-WormBase and 0.88 for GG-WormBase comparison) (Fig. 1A). TransRate (Smith-Unna et al. 2016) transcriptome overall assembly scores were 0.39 for GF and GG, well placed within the 90th percentile of scores of other assemblies using different assembly algorithms and species (Fig. 1B; Smith-Unna et al. 2016). Predicted ORFs encompassed 96.4% (GF) and 95.9% (GG) of the 3131 universal single-copy orthologs of nematode Benchmarking Universal Single-Copy Orthologs (BUSCO) gene models (odb10) (Simão et al. 2015) in full length (Fig. 1C), showing the comprehensiveness of the assembly. Indeed, the predictions cover 18,794 (93.4%—GF) and 18,858 (93.7%—GG) of the 20,127 *C. elegans* WormBase gene models (Fig. 1D; Supplemental Table S2), with 73.8% (GF) and 70.0% (GG) predictions having high sequence coverage (80%–100%) with their respective WormBase gene model (Fig. 1D). All benchmarks showcase that our approach results in comprehensive annotations with mostly complete and precise models. In all quality measures, the GF assembly mode performs better than the GG approach.

Machine-learning-based algorithm to judge gene model accuracy

Although most of the assembled transcripts were indeed full length compared with the current WormBase annotation of *C. elegans*, our assembled transcriptomes still included some transcripts that were only partially assembled (Fig. 1D). The issue of transcript fragmentation in assemblies from RNA sequencing data is a well-known problem and has been the focus of many studies (Treangen and Salzberg 2011). These artifacts are typically caused by low read coverage at a locus, repetitive regions, differential expression of different exons, polymorphism, and sequencing errors, which might potentially lead to local assembly errors (Treangen and Salzberg 2011). In our case, most of the partially assembled transcripts are WormBase genes that were split during the assembly process and thus are represented as separate nonoverlapping transcripts (Fig. 2A; Supplemental Fig. S3; Supplemental Table S3). This fragmentation issue is much more evident in the GG assembly approach. As including such fragmented contigs in downstream analyses can cause misinterpretation, we aimed to identify incomplete transcripts also when no comparison to an existing well-curated annotation is possible. To address this, we applied supervised machine learning using random forest (RF). The algorithm was trained using the *C. elegans* assemblies with different transcript-specific input features retrieved from Trinity (Grabherr et al. 2011), TransDecoder (Bryant et al. 2017), and TransRate (Supplemental Table S4; Smith-Unna et al. 2016). The completeness of the transcript was assessed by comparing the predicted ORF to the respective WormBase protein annotation. Because the underlying assembly algorithms of the GF and the GG approaches are different, we generated independent prediction models for GF and GG. When comparing the predicted to the observed WormBase annotation–based completeness, the Pearson's correlation was 0.96 for GF and 0.88 for GG. The most decisive features for the prediction model were the length of the ORF and the full transcript for GF, and the full transcript length and expression level (transcripts per million [tpm]) for GG (Fig. 2B; Supplemental Table S4). As we aimed to predict transcript coverage for different nematode species, we evaluated the performance of our machine-learning models using more phylogenetic distant, but well-annotated species. For benchmarking, we assembled our own gene models with publicly available RNA sequencing data of the nematode *C. briggsae*, the fruit fly *Drosophila melanogaster*, the green land plant *Arabidopsis thaliana*, and the human H1-hESC cell line (Supplemental Table S5) using the same assembly workflows as applied in *C. elegans*. The *C. elegans* trained predictors showed very high accuracy in determining the transcript completeness in all four species (Fig. 2C; Supplemental Table S5). This strongly indicates that our gene model predictors should be applicable to a broad range of species even beyond nematodes, thus enabling efficient filtering of fragmented contigs across diverse transcriptome assemblies.

Further gene model refinement by applying proteotranscriptomics

To provide experimental evidence for the predicted ORFs at the protein level, we measured the proteome of the same *C. elegans* mixed-stage sample by high-resolution mass spectrometry. We used either the WormBase, GF, or GG predicted ORFs as a protein sequence database to associate the roughly 1.6 million MS/MS

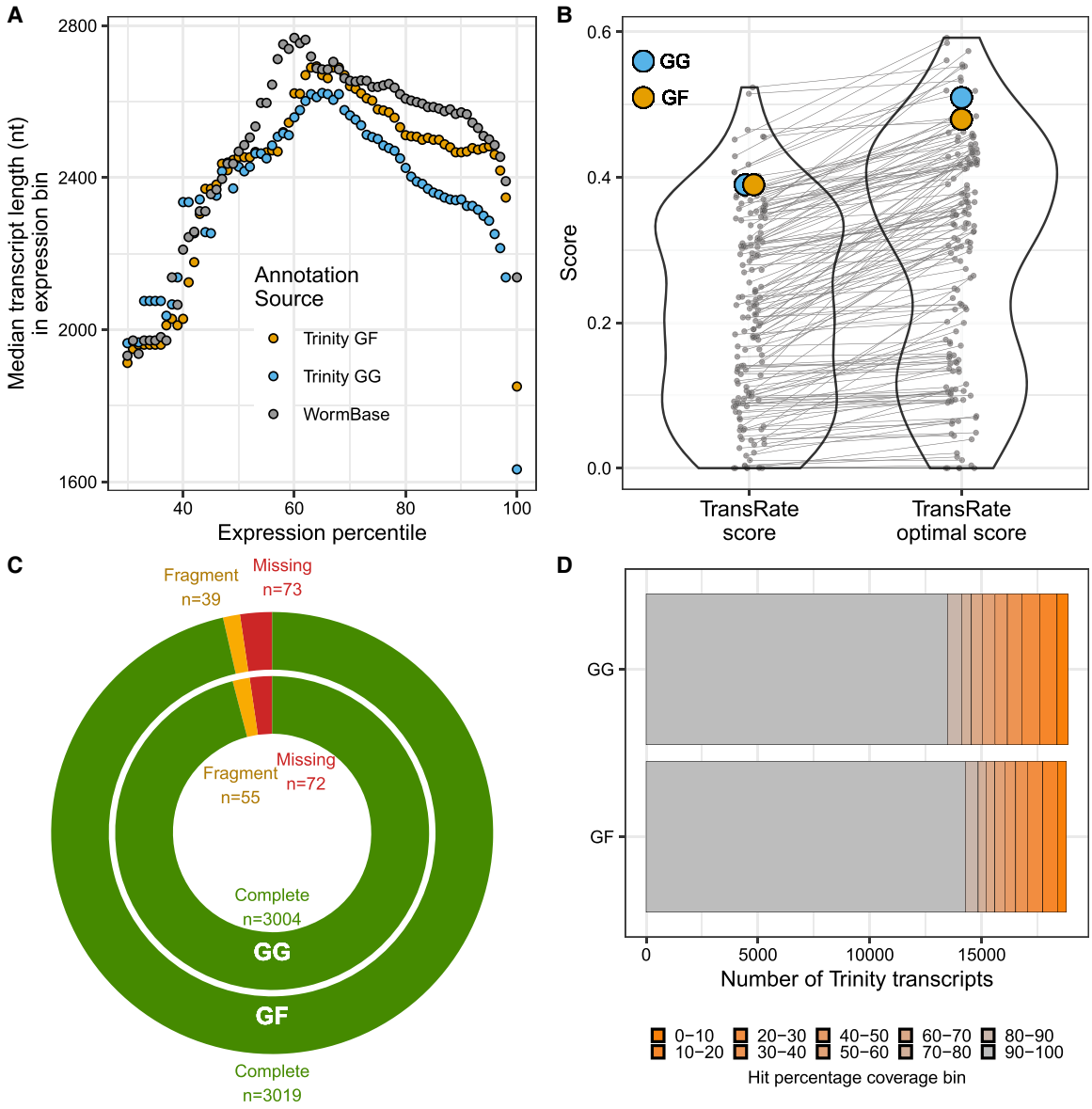
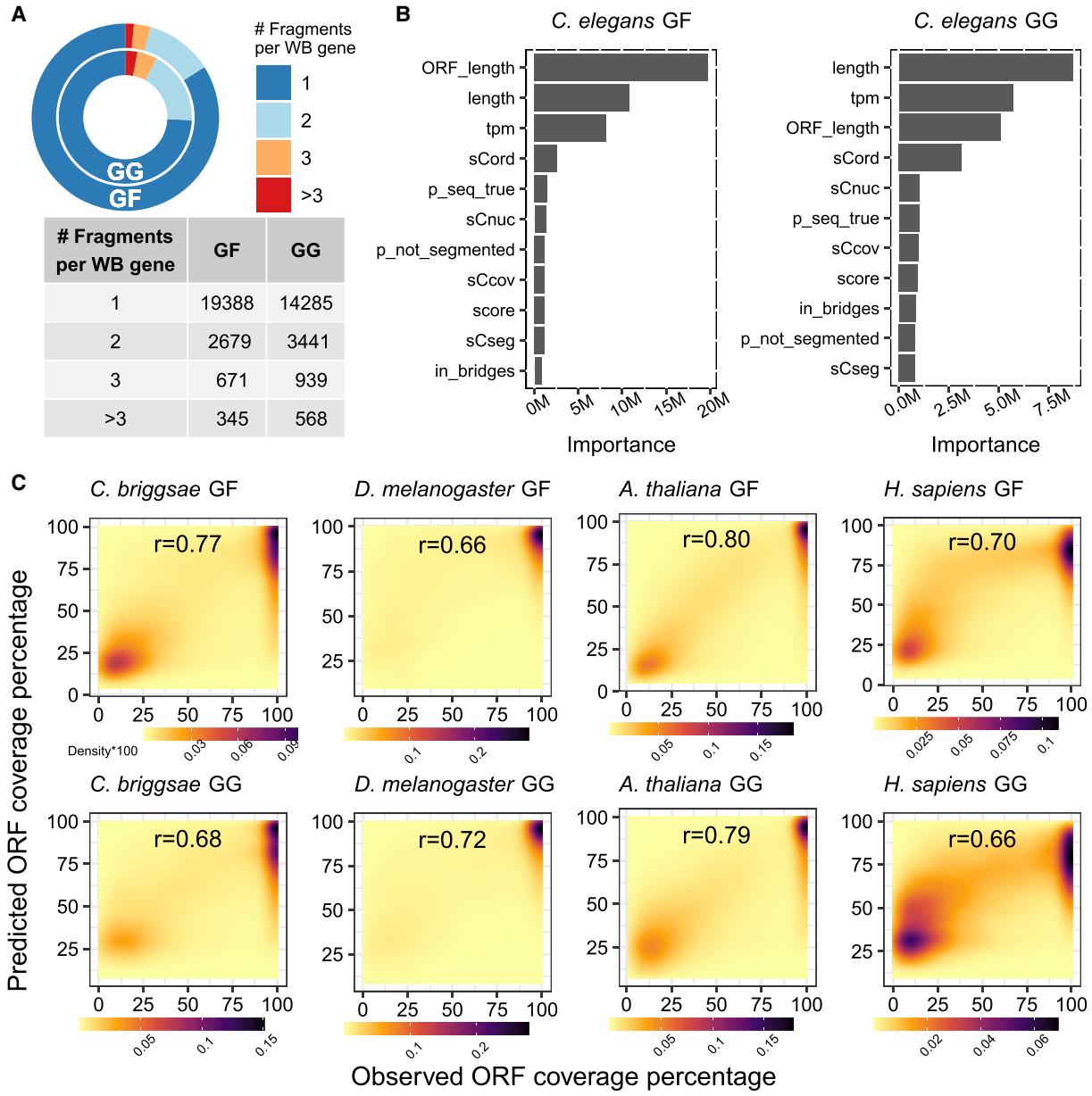


Figure 1. Benchmarking transcriptome assembly in *C. elegans*. (*A*) Median length of transcripts (N50) across all expression bins (ExN50) for the genome-free (GF) and genome-guided (GG) assembly compared with the WormBase annotation. (*B*) TransRate scores and TransRate optimal scores of the *C. elegans* GF and GG transcriptome assembly compared with other assemblies of different species and various assembly algorithms (Smith-Unna et al. 2016). (*C*) BUSCO analysis statistics of the GF and GG transcriptome assembly. (*D*) Bar chart summarizing transcript sequence coverage comparing GF and GG transcripts to the corresponding WormBase annotation. Completeness (hit percentage coverage) was determined by percentage overlap between the predicted transcript sequence with its equivalent WormBase annotation.

spectra with tryptic peptides from these annotations. Trinity assemblies showed a comparable amount of identified peptides compared with the WormBase annotation (97% for GF and 98% for GG) (Fig. 3A). All three assemblies showed comparable numbers of protein identification exceeding 7000 proteins (Fig. 3B). We observed that ORFs with peptide evidence are highly enriched for full-length WormBase transcripts (Fig. 3C). To prevent fragmented proteins in our proteotranscriptome assemblies even more efficiently, we additionally filtered out any ORF with a predicted completeness level <80% as judged by our machine-learning algorithm. After this filtering, the identified proteins from the GF and GG assemblies include 95% and 93% of the identified proteins from WormBase, respectively (Fig. 3D).

Furthermore, we found 839 short proteins (fewer than 100-amino-acid length) in the GF and 830 in the GG assembly with predicted completeness levels >80% that are supported by at least two peptides, at least one of them being unique (Supplemental Table S6). Comparing these proteins to the *C. elegans* database of small proteins from SmProt (Hao et al. 2018), we identified BLASTP hits with known short proteins for 161 predictions (19%) in GF and 96 (12%) in GG (Supplemental Table S6). As the SmProt database consists of predicted small proteins from ribosome profiling data, these results emphasize the high sensitivity and reliability of our approach, confirming some of the SmProt predictions but also providing strong evidence for hundreds of additional *C. elegans* small proteins.



# Fragments per WB gene	GF	GG
1	19388	14285
2	2679	3441
3	671	939
>3	345	568

Figure 2. Benchmarking machine-learning-assisted filtering of fragmented transcripts in assemblies. (*A*) Pie chart and table representing the fragmentation status of the *C. elegans* GF and the GG transcriptome assembly. Shown are the proportions of WormBase genes overlapping with Trinity assembled transcripts. (*B*) Bar chart depicts contributions of individual features to the random forest models of GF and GG. A detailed description of features and their sources is provided in Supplemental Table S4. (*C*) 2D kernel density plots of the observed versus predicted completeness using the GF and GG model established in *C. elegans* in four other model organisms.

Although our approach missed 305 annotated WormBase proteins (<5% of the detected proteome), we found two predicted proteins with strong peptide evidence in GF and GG that were not reported in previous *C. elegans* annotations of WormBase. For these two genes, we could detect dynamic expression at the RNA level using previously published developmental transcriptomic time courses of *C. elegans* (Supplemental Fig. S4; Boeck et al. 2016; Levin et al. 2016). The first protein (to be included as F54D10.10 in WormBase release WS286), whose transcript sequence maps to Chromosome 2, has a length of 138 aa and is supported by three unique peptides and an overall mRNA level of 18 tpm (83rd percentile) (Fig. 3E). We found an expression peak of F54D10.10 in early embryonic stages (90 min after the fourth division of the AB cell) in both data sets (Supplemental Fig. S4A–C). Although there were no homologs in WormBase, in NCBI we found a predicted protein from *C. remanei* (hypothetical protein GCK72_007074), albeit it only shows 39% sequence identity (Supplemental Material). The transcript sequence of the other protein (to be included as Y34B4A.20) maps to Chromosome X, has a length of 155 aa, is supported by four unique peptides, and showed an overall mRNA level of 12 tpm (80th percentile) (Fig. 3F). Although again there were no homologs in WormBase, we found predicted proteins for *C. remanei, C. briggsae, C. japonica,* and *Caenorhabditis nigoni* in NCBI (Supplemental Material).

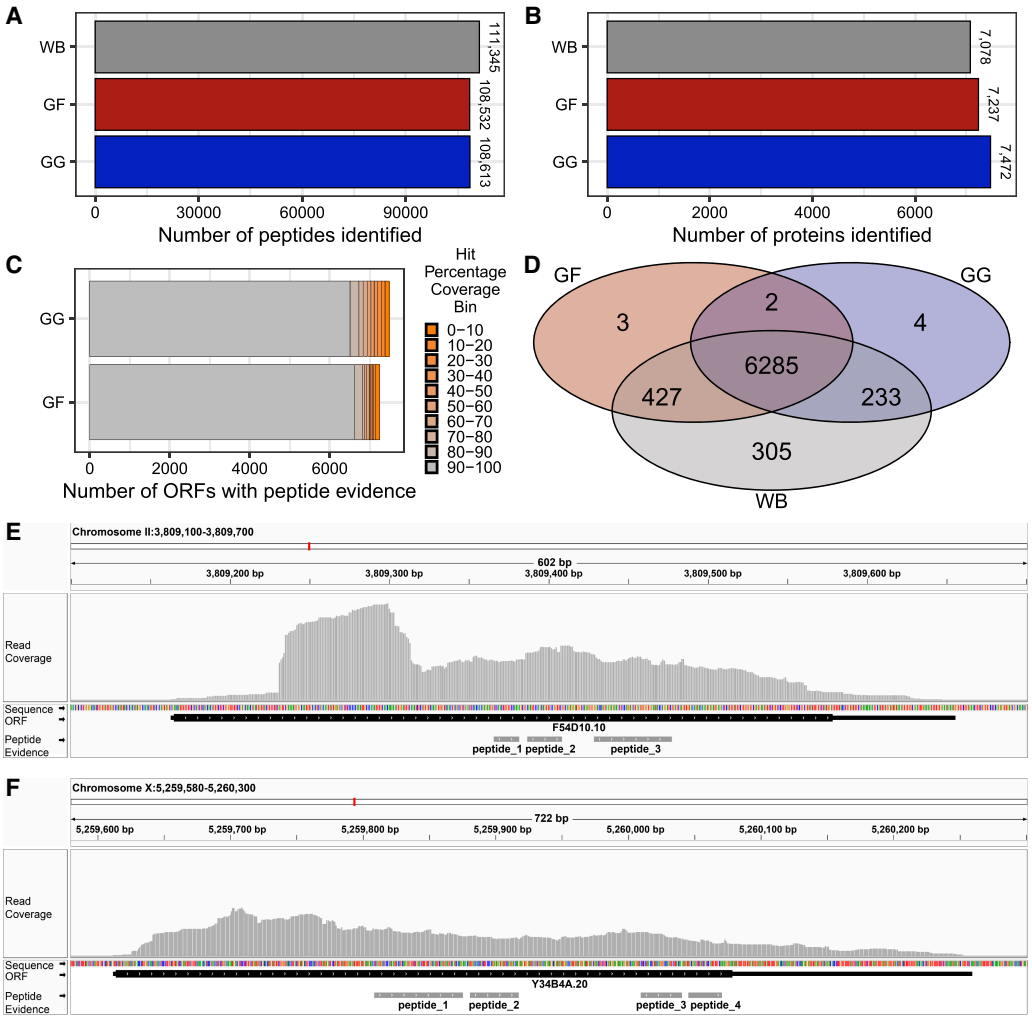


Figure 3. Proteotranscriptomics in *C. elegans*. (*A*) Number of peptides identified in the *C. elegans* WormBase, GF, and GG assemblies. (*B*) Number of individual proteins with peptide evidence for WormBase, GF, and GG proteomes. (*C*) Stacked bar plot of all ORFs of the GF and GG assembly with peptide evidence grouped by the level of coverage with the respective WormBase entry. (*D*) Venn diagram depicting the overlap between the identified proteins using WormBase, GF, or GG assembly as search space for peptide identification. (*E*) Visualization of the new *C. elegans* gene *F54D10.10* via the Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al. 2013) aligned to the *C. elegans* genome sequence. Presented are read coverage (gray peak track), ORF structure (black bar; thick bar represents translated region), and position of peptide evidences (gray bars). (*F*) Same as *E* for *Y34B4A.20*.

Y34B4A.20 seems to be expressed exclusively in larval stages (Supplemental Fig. S4D,E).

Combined, benchmarking in *C. elegans* shows that our approach can recapitulate most of the current annotations of this very comprehensively studied species but also facilitate the detection of nonreported coding genes.

Gene annotation for additional *Caenorhabditis* species and for phylogenetically distant and non-genome-sequenced species

Having validated that our proteotranscriptomics approach yielded outstanding results for *C. elegans* in the GG as well as the GF mode, we went on to perform transcriptome assemblies for five additional *Caenorhabditis* species with available genomes and gene annotations. Although *C. elegans* and *C. briggsae* have fairly well-evidenced annotations, most of the other species lack experimental validation, possessing mostly predicted ORFs (Supplemental Fig. S2). For the assembly of the additional *Caenorhabditis* species, we achieved similar high-performance measures in terms of TransRate scores (Supplemental Table S1), BUSCO benchmarks (Fig. 4A, see *Caenorhabditis* panel), and number of identified peptides (Fig. 4B, see *Caenorhabditis* panel).

However, although for the well-annotated species such as *C. elegans*, *C. inopinata*, and *C. briggsae* the WormBase annotation allowed for slightly better protein identification compared with our own ORF predictions (ranging from 1.3% to 2.5% better), for *C. japonica*, *C. remanei*, and *C. brenneri*, we observed significant improvement with our new assemblies (identification increases ranging from 5.9% to 14.9%) (Fig. 4C, see *Caenorhabditis* panel). Performance of the GF mode was slightly better for many species (improvements of 0.4% to 14.9%). This outlines the strength of the GF approach, especially for species with less well-assembled genomes.

With these quality confirmations in the additional *Caenorhabditis* species, we confidently took the same approach and expanded our annotation to nematode species outside the

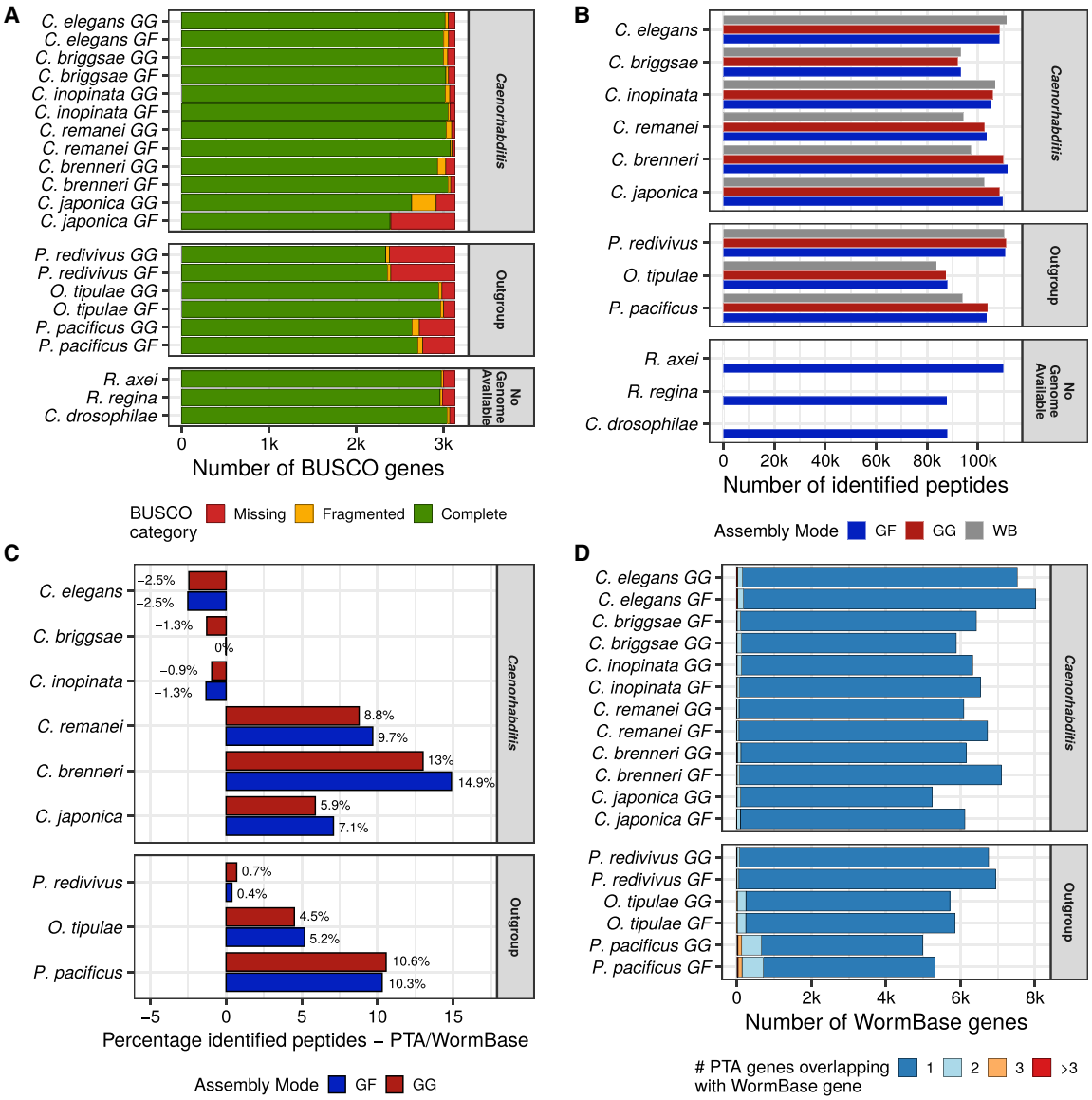


Figure 4. Proteotranscriptomics assembly (PTA) of 12 nematode species. (*A*) BUSCO metrics for all species and the two assembly modes, GF and GG. (*B*) Bar plot of mass spectrometry–identified peptides belonging to protein entries of the respective annotation source (GF, GG, and WB) for each species. (*C*) Peptide identification improvement of GF and GG annotations compared with WormBase annotations. (*D*) Number of GF and GG proteins with peptide evidence overlapping with the respective WormBase protein annotation for each species. Light blue, orange, and red groups represent WormBase entries that are covered by more than one proteotranscriptomics-validated protein.

Caenorhabditis lineage with available genome assemblies such as *P. pacificus*, *Oscheius tipulae*, and *Panagrellus redivivus* ("outgroup" panel) but also to species without a sequenced genome such as *Rhabditoides regina*, *Rhabditella axei*, and *Caenorhabditis drosophilae* ("no genome available" panel). These species span an evolutionary distance of 22 million years (Weinstein et al. 2019) and constitute a highly interesting set for further phylogenetic analysis. The TransRate scores of the assemblies of all species were exceptionally high, even higher than for the previous *Caenorhabditis* species (Supplemental Table S1). BUSCO comprehensiveness is high across most species regardless of assembly mode (GF and GG), whereas the GF approach again achieved slightly better representation (Fig. 4A). It is noteworthy that *P. pacificus* and *P. redivivus* transcriptome assemblies showed a lower representation of BUSCO genes, and their lack was independent of the generated transcriptome assemblies as it can also be observed in the respective WormBase annotations (Supplemental Fig. S5). The overlap between the missed BUSCO genes in these two species is highly significant (more than 250 ORFs, P-value $< 10^{-75}$, hypergeometric test), arguing for a strong sequence divergence or loss of these genes in more distant nematode lineages. Using the different assemblies for peptide identification in proteomics, we observed similar identification levels as for the *Caenorhabditis* group (Fig. 4B). Without exception, outgroup species showed improved identifications compared with their WormBase annotations (0.5% *P. redivivus*, 4.9% *O. tipulae*, 10.5% *P. pacificus* increase on average) (Fig. 4C), emphasizing the ability of our approach to improve annotations beyond state-of-the-art methods.

The proportion of genes with fragmentation in the assemblies is very low (ranging between 1% and 5%) (Fig. 4D), except for *P. pacificus* with exceptionally high levels of presumably split genes. This observation can have two causes: either our approach assembled more fragmented ORFs for *P. pacificus* or the current *P. pacificus* annotation from WormBase includes mistakenly merged ORFs. The fusion of genes is normally a very rare event in evolution, thus wrong prediction by automated genome annotations is the more plausible cause (Melsted et al. 2017; Levin et al. 2020). To check whether this is the case, we compared the *P. pacificus* WormBase gene models to the well-established *C. elegans* WormBase gene models in order to detect incoherence in ortholog lengths. We indeed detected significantly reduced ortholog coverage in *P. pacificus* proteins that were covered by more than one of our Trinity transcripts. These might represent incorrectly merged genes. In 88.8% (893 of 1006) of our predicted ORFs that were shorter than the *P. pacificus* WormBase annotation, we found that the *C. elegans* models indeed fit the shorter reading frame (Supplemental Fig. S6A,B). In addition, applying our machine-learning-based completeness prediction, some of the *P. pacificus* WormBase proteins were flagged for artifactual fusions. The corresponding Trinity-predicted proteins showed high machine-learning-predicted completeness levels while overlapping only partially with the *P. pacificus* WormBase orthologs (Supplemental Fig. S6C). In agreement with this, recent studies have indeed reported that some of the initial *P. pacificus* protein annotations were false merges of individual genes (Rödelsperger et al. 2019; Rödelsperger 2021). Although 9% (64 proteins) of our predicted fusion artifacts were reported in these two studies, we provide evidence for 641 additional cases (Supplemental Table S7). These findings support that we were able to refine ORFs that were likely falsely merged in former annotations.

The predicted ORFs with peptide evidence from the GF and GG assemblies of all nine species with an annotated genome show a very high overlap with WormBase (Supplemental Fig. S7). As expected from a well-curated model species, for *C. elegans*, we found the lowest number of proteins that were exclusively detected in our assemblies, but also missed relatively few WormBase proteins by our approach. The remaining species can be divided into two categories: (1) species showing a moderate number of not yet annotated proteins with fair amounts of missed WormBase proteins—*C. briggsae*, *C. inopinata*, *P. redivivus*, and *O. tipulae*, and (2) species with high numbers of not yet annotated proteins and strongly increased numbers of missed WormBase proteins—*C. brenneri*, *C. japonica*, *P. pacificus*, and *C. remanei*. Thus, some species are already quite well annotated, whereas in others, we can provide more improvements.

Applying the same methodology that delivered solid benchmarking results in *C. elegans* to 11 additional species, we observed highly consistent performance, enabling the annotation of at least 6300 ORFs with peptide evidence in each species (Supplemental Table S2).

Insights into nematode evolution with a consolidated phylogeny

Here, we improved annotations for nine nematode species and provide the first high-coverage annotation for three additional nematode species. This set of species allows for interesting evolutionary analyses as they encompass seven species of the *Caenorhabditis* genus, two species of the extended group of Eurhabditis (*R. axei* and *O. tipulae*), and three outgroup species still belonging to the order of Rhabditida (*P. redivivus*, *P. pacificus*, and *R. regina*).

We first determined orthology groups for all predicted ORFs with >80% completeness levels using the orthology detection program ProteinOrtho (Lechner et al. 2011), resulting in 23,090 orthology groups that contain orthologs in at least two species; 3261 groups (14%) have orthologs across all 12 species (Fig. 5A). As expected, these orthologs have a highly significant overlap with the nematode BUSCO set (P-value < 10^{-337}, hypergeometric test) and are enriched with knockout phenotypes related to fertility and embryonic development (Supplemental Table S8). These findings emphasize the importance of these core genes in the highly conserved developmental program as has already been reported by others (Davidson and Erwin 2006; Kalinka et al. 2010; Levin et al. 2016; Malik et al. 2017).

Another interesting group are orthologs that were only detected in the *Caenorhabditis* genus (568 orthology groups). These proteins are enriched with various knockout phenotypes and Gene Ontology terms reflecting functions in cell division (Supplemental Table S8). Unique features of early embryonic cell divisions such as asymmetry and spindle oscillation have been shown to have emerged uniquely within *Caenorhabditis* (Delattre and Goehring 2021). We also found functional enrichments for processes involving the addition and removal of phosphate groups, especially on serine and threonine residues. Many of these kinases and phosphatases were shown to be involved in cell division regulation (Nasa and Kettenbach 2018), and hence, their unique presence in *Caenorhabditis* could be the basis of the *Caenorhabditis*-specific cell-cycle mechanisms. We could substantiate these results using STRING database (STRINGdb) associations, which enables the identification of protein–protein interaction networks and functional enrichment analysis. Interrogating the list of *Caenorhabditis*-specific ORFs, STRINGdb generates two main clusters enriched with the terms "cell division" and "phosphorylation/dephosphorylation," which are even interconnected (Supplemental Fig. S8; Supplemental Table S8).

For 357 orthology groups specific to the two Eurhabditis species *R. axei* and *O. tipulae* not existing in *Caenorhabditis* and another 48 orthology groups restricted to the non-Eurhabditis species (*P. redivivus*, *P. pacificus*, and *R. regina*), we did not find obvious functional enrichments (for all orthology groups, see Supplemental Table S9).

To enable rigorous evolutionary comparisons and to construct a phylogeny based on thousands of genes, we restricted our orthology analysis to the proteotranscriptomics-validated ORFs, that is, those supported by peptide evidence. Thus, we used 1516 orthology protein groups that only have one-to-one orthologs across all species. By multiple alignment of these protein sequences, we reconstructed individual gene trees for each orthology group with three different methodologies, selected the best scoring tree, and finally combined the individual gene trees into a phylogenetic species tree (Fig. 5B). The topology of this tree is in accordance with the recently published taxonomic relationship between nematodes (Ahmed et al. 2021), but we substantiate the phylogeny with an extensive set of one-to-one orthologs. Hereby, our methodology of combining de novo assembly of transcriptome and the integration of peptide evidence facilitated a comprehensive phylogenetic analysis.

Signatures of molecular evolution

Using one-to-one orthologs supported by proteotranscriptomics, we set out to estimate molecular evolution across Rhabditida. We

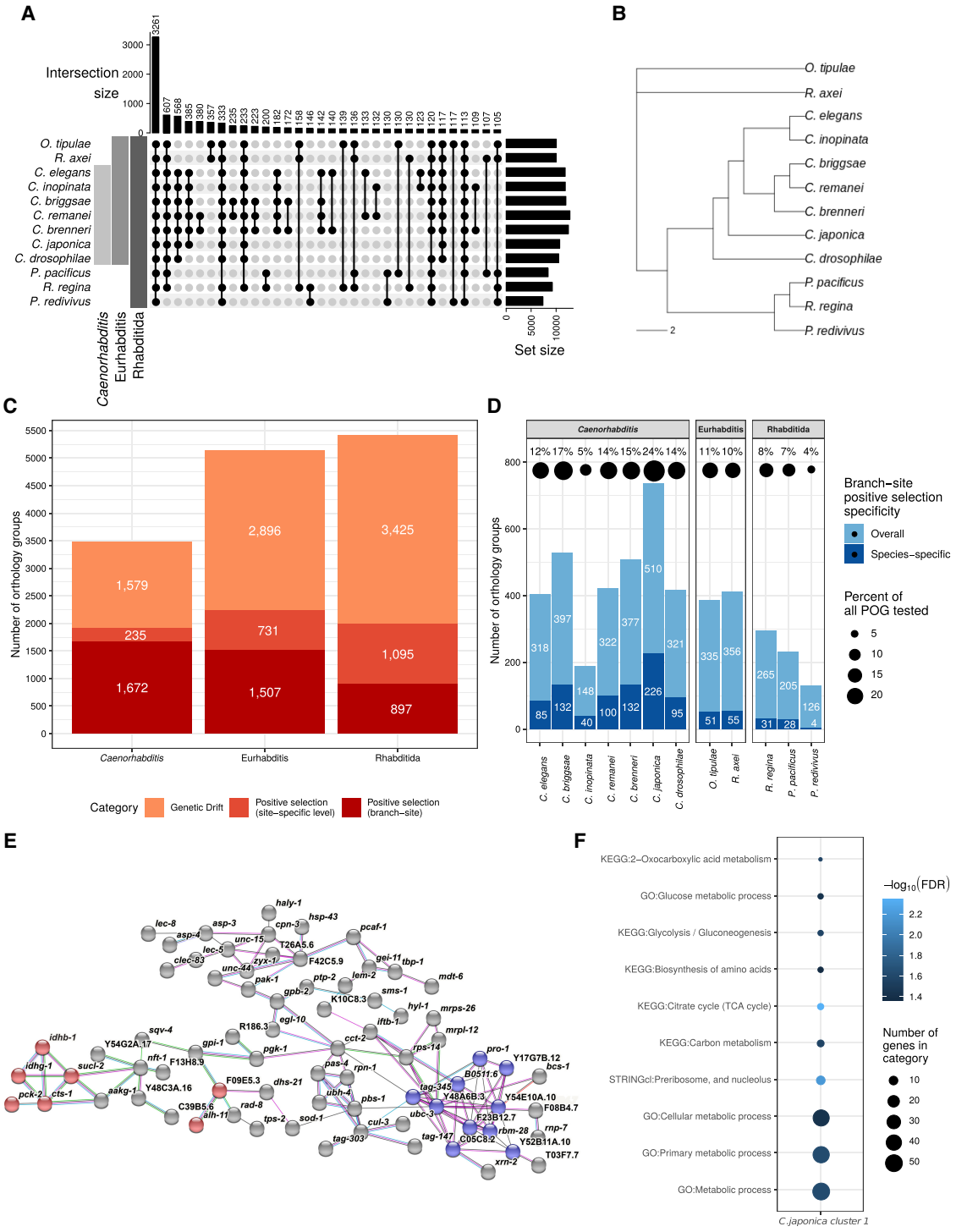


Figure 5. Orthology and phylogenetic relationships. (*A*) Upset plot depicting the number of orthology groups shared between different species. (*B*) Combined unrooted phylogenetic tree establishing the relationships between all studied species. The tree is based on individual gene trees of 1516 orthology groups that contain exactly one orthologous gene for each of the 12 studied species. The branch length is defined as the number of amino acid substitutions per site. (*C*) Distribution of genetic drift and positive selection in orthologous groups encompassing *Caenorhabditis*, Eurhabditis, or Rhabditida. Positive selection is reported separately for detection either in the site-specific (light red) or branch-site (dark red) analysis. (*D*) Distribution of orthology groups with significant signatures of branch-site-specific positive selection across species. ProteinOrtho groups (POGs) are colored for positive selection either in one (dark blue) or multiple (light blue) species. The percentage of species-specific positive selection instances (dark blue) among all POGs that contain orthologs from the respective species are shown on *top* of the bars. (*E*) STRINGdb network of *C. japonica* proteins with positive selection signals. Nodes represent single proteins, and edges represent protein–protein associations provided by STRINGdb. Edge colors represent protein–protein association types: blue, from curated databases; pink, experimentally determined; green, gene neighborhood; red, gene fusions; dark blue, gene co-occurrence; black, coexpression; and purple, protein homology. Proteins belonging to the glycolysis and TCA cycle network are marked in red; proteins of the ribosome biogenesis cluster are colored in blue. (*F*) KEGG, Gene Ontology, and STRINGdb cluster terms enriched in the protein cluster depicted in *E*.

determined d_N/d_S across the orthology groups and subsequently scored signals of positive selection. Adaptive evolution is not easily identifiable in a phylogeny-encompassing species with large evolutionary distances, and indeed, we found overall d_N/d_S values to be higher for orthology groups including only *Caenorhabditis* or Eurhabditis species (Supplemental Fig. S9A). Hence, we divided our analysis into three sets taking phylogenetic distance into account and focused on orthologous proteins shared between at least three of either *Caenorhabditis*, Eurhabditis, or Rhabditida to investigate different evolutionary models (M0, M1, M3, M7, and M8) using 3486, 5134, and 5417 orthology groups, respectively. Using multiple sequence alignments of all proteins from the different groups and the M7 and M8 site models, we were able to identify genes that were under positive selection across the studied nematode species. As expected, the majority of genes show genetic drift: 45.3% for *Caenorhabditis*, 56.4% for Eurhabditis, and 63.2% for Rhabditida (Fig. 5C). As we were mostly interested in signals of positive (adaptive) evolution, we evaluated branch-site models for those genes that had positive selection signals at the site-specific level to identify branches under selection. By this stepwise analysis, we found branch-site-specific positive selection signals in 1672 orthology groups (47.9%) for *Caenorhabditis*, 1507 (29.3%) for Eurhabditis, and 897 (16.5%) for Rhabditida (Fig. 5C). To rule out biases in assembly efficiency that derive from overall genetic diversity (e.g., different levels of heterozygosity) (Romiguier et al. 2012) owing to different reproduction modes, we compared the resulting number of high-quality assembled transcripts across species that reproduce primarily by selfing (androdioecious) or mating (gonochoristic) and found no significant difference (Supplemental Fig. S9B). We also did not detect any significant reproductive mode–dependent trends in the terminal branch average d_N/d_S values (Supplemental Fig. S9C), showing that we do not observe such biases in our data.

When evaluating the composition of the ProteinOrtho groups with positive selection signals at the branch level in the respective species, we found these signals to distribute unevenly across species of the subsets (Fig. 5D). As the species are themselves unequally represented in the orthology groups, we normalized the number of signals by the number of orthology groups for each species. When comparing these proportions across the different subsets, we found, as anticipated, that signals of positive selection diminish with increasing phylogenetic distance between the species (Supplemental Fig. S9D). We observed that some species show proportionately more genes under positive selection (Fig. 5D; Supplemental Fig. S9D). This also applies when selecting genes that show branch-site-positive signals in only one of the species of the subgroups (*Caenorhabditis*, Eurhabditis, or Rhabditida) (Fig. 5D, dark blue).

In the *Caenorhabditis* genus, *C. japonica* shows the highest proportion of positively selected genes, reaching 24% (226 genes). To functionally characterize these genes, we used STRINGdb (Supplemental Table S10; Szklarczyk et al. 2021). We found five clusters of proteins with associations. The biggest cluster consists of 70 genes highly enriched with metabolic functions mainly in the glycolysis and the tricarboxylic acid (TCA) cycle (Fig. 5E,F). Whereas *C. elegans* is a free-living nematode, *C. japonica* has a species-specific phoretic relationship with the hemipteran *Parastrachia japonensis* (Tanaka et al. 2012). Under unfavorable conditions, *C. elegans* forms a long-lived larva (dauer larva) characterized by reduced metabolic activity, elevated superoxide dismutase expression increasing resistance to oxidative stress (Larsen 1993), and increased expression of several heat shock proteins (Dalley and Golomb 1992) that can survive up to 3 mo. In contrast, attached to *P. japonensis*, the *C. japonica* stress-resistant dauer stage lasts naturally for ~11 mo while waiting for yearly fruit ripening. This is more than three times longer than has been observed for *C. elegans* dauer larvae. Previous studies have shown that when *C. japonica* were moved to laboratory conditions, the longevity of their dauer larvae shortens to only 10 d (Tanaka et al. 2012). As the dauer developmental switch is accompanied by a switch in the metabolic pathways from aerobic to anaerobic processes, the identified metabolic gene enrichment suggests that *C. japonica* adapted some of the genes involved in the aerobic pathways or in the switch between the two pathways. In *C. elegans* dauer larvae, citric acid cycle activity is reduced relative to that of the glyoxylate cycle, consistent with utilization of stored lipids (Wadsworth and Riddle 1989; O'Riordan and Burnell 1990). As the gene cluster includes several enzymes involved in the TCA pathway (Fig. 5E, red; Supplemental Fig. S10), this might reflect released selective pressures that facilitated the coadaptation of *C. japonica* with its host *P. japonensis*.

Within the same cluster of *C. japonica* genes with positive selection signals, an additional tight subnetwork of genes involved in ribosome biogenesis was detected (Fig. 5E, blue). As ribosome biogenesis is a basic process that has broad effects, we could not pinpoint specific physiological features connecting this process to the ecology or biology of *C. japonica*. Nevertheless, common knockout phenotypes of these genes are related to slow growth and larva viability. In general, we found ribosomal proteins under positive selection in the nematodes *C. briggsae*, *C. drosophilae*, *C. inopinata*, *O. tipulae*, *P. pacificus*, *R. axei*, and *R. regina* (Supplemental Table S10). This suggests that changes at ribosomal complexes might be an important evolutionary toolbox for environmental adaptation. Such selective dynamics have been very recently reported for yeast species (Sultanov and Hochwagen 2022), but this phenomenon has not been described in nematodes yet.

Furthermore, we also found a large cluster of 18 proteins with positive selection in *C. brenneri* enriched for functions involved in fatty acid metabolic processes (Supplemental Table S10). Previous work showed that ascaroside signaling is widely conserved among nematodes, and many basic components are produced by a large diversity of species (Choe et al. 2012). However, of the nine *Caenorhabditis* species that were analyzed, all except *C. brenneri* were found to produce indole ascarosides (Choe et al. 2012). The diversity of biological functions regulated by ascarosides is paralleled by their structural diversity, which depends primarily on the variability of their aglycones, which in turn originates from the co-option of a primary metabolic pathway, the peroxisomal β-oxidation of fatty acids, in ascaroside biosynthesis. As many of the enzymes in the *C. brenneri*–positive selection cluster have functions in fatty acid α- or β-oxidation, this might explain the divergence of environmental signaling in this species.

The described examples showcase the evolutionary relevance that can be obtained from selective signals using validated protein-coding transcriptome information within an extended nematode phylogeny.

Discussion

By combining readily available short-read RNA sequencing with high-resolution mass spectrometry–based proteomics in an approach called proteotranscriptomics, we provide and interrogate extensive protein-coding gene annotations for 12 nematodes, including species with low-quality or nonexistent genome

assemblies. By implementing a novel machine-learning approach, we are able to detect incomplete transcript assemblies and remove such artifacts. Benchmarking the annotation efforts by comparison to the bona fide *C. elegans* proteome, we show that the approach performs very well, recapitulating 94% of the proteins that can be detected by mass spectrometry in our experimental setup. Furthermore, we present two genes (*F54D10.10* and *Y34B4A.20*) that have not been reported in prior *C. elegans* annotations, emphasizing the power of our method.

Although the precision of the method was very high, the restrained proteome coverage of the mass spectrometry measurements poses a certain limit to the comprehensiveness of our annotation. In principle, the overall range of detected proteins could be extended by applying technological and methodological adjustments (Levin and Butter 2022). However, even with RNA sequencing, we could detect meaningful counts for only ∼43% of the annotated *C. elegans* genes (8581 transcripts with at least 10 detected tpm). Indeed, we were able to detect peptide evidence for ∼87% of the genes that are transcribed in our samples, including all developmental stages. For these genes, we see high correlations between transcript expression level and protein intensity but also peptide sequence coverage (Supplemental Fig. S12). Despite these coverage restrictions, we show that the resulting data are solid and enable findings that would be impossible, especially for species that have no assembled genome yet. Although we are not able to assemble all potential ORFs, we are confident that the scope of our analyses actually benefits from the extra layer of confidence that the ORFs under investigation are actively expressed proteins. Applying the proteotranscriptomics approach to more species, we not only improve annotations of nine species but also provide annotations for three additional currently nonannotated species.

The presented data are a valuable genetic resource for the scientific community, as they facilitate research in a larger group of diverse nematode species for future transcriptomic, proteomic, and comparative evolution studies. The GF mode, which does not depend on genome assembly, shows superior results reflected in less fragmentation, more peptide and protein identifications, and better BUSCO coverage in most cases. It therefore seems the method of choice and is universally applicable even for non-genome-sequenced species. The better performance of GF is intuitive for species with highly fragmented genome assemblies, as high numbers of gaps hamper precise transcript assembly when relevant reads do not map to the genome and thus are excluded from the GG assembly process. For species with high-quality genome assemblies such as *C. elegans*, the interpretation of this result is not as straightforward, as here we would expect the GG approach to work better than GF. However, a few technical aspects of the GG assembly process might explain our observations. In the Trinity GG approach, aligned reads are clustered into coverage groups based on the alignment. Then each read cluster is assembled using the standard Trinity de novo assembly. Although this approach makes the assembly more straightforward in terms of computational complexity, it bears a few pitfalls. First, to avoid assembling potentially wrong transcripts containing very long artificial introns, the algorithm applies a threshold of maximum intron length within the read alignments. For all evaluated species, this threshold was set to 3500 bases based on previous reports (Wu et al. 2013). Despite this threshold being important to avoid the assembly of potentially wrong transcripts, it will prevent the full-length assembly of any transcript that genuinely has longer introns (0.6% of all introns). Indeed, in *C. elegans* we observe that 39% of the GG assembled transcripts that show higher fragmentation compared with the GF approach have introns that are longer than 3500 bases (Supplemental Fig. S11A). Another important limiting factor of the GG approach is the read coverage of a locus. Loci with low coverage have higher chances to be fragmented, as there will be only very few reads connecting read coverage groups across the locus. Indeed, we observe that transcripts with higher fragmentation compared with the GF approach have significantly lower read coverage (tpm) than transcripts that were fully assembled in both the GF and GG approaches (Supplemental Fig. S11B). This property can also be extracted from the feature importance measures of our machine-learning model (Fig. 2B). In the GG model "tpm" is much more relevant to the completeness prediction than in the GF model. The machine-learning filtering we introduced does indeed filter out 77% of these fragmented transcripts (Supplemental Fig. S11C).

Dissecting homology relationships among the genes in these 12 species at the transcriptome level, we could predict over 23,000 orthologous families across the different species. These include orthology groups that have not been described before, for example, one group encompassing orthologs across all species except *C. elegans*, *C. briggsae*, and *P. pacificus* (group ID 7609) (Supplemental Table S9). One of the genes in this group is the predicted *P. redivivus* gene Pan_g7772.t1. We found unreported orthologs for the other eight species. This emphasizes the opportunities of our approach, which is independent of previous annotations, as opposed to many gene prediction pipelines that heavily rely on comparison to model species as reference, causing newly evolved or lost ancient genes to be missed.

Characterizing the orthology groups unique for *Caenorhabditis*, we found enrichments of networks related to cell division and spindle organization. Although it is known that species of the *Caenorhabditis* genus have unique spindle formation mechanisms (Delattre and Goehring 2021), the assembly and disassembly of the required protein complexes are still not fully understood. We found several genes within the *Caenorhabditis*-specific genes that were suggested to be involved in this process: *spd-2* and *spd-5* (Hamill et al. 2002; Woodruff et al. 2014; Conduit et al. 2015; Magescas et al. 2019; Stenzel et al. 2021), *rod-1* (Henen et al. 2021), *klp-19* (Bayliss et al. 2003; Schlaitz et al. 2007; Müller-Reichert et al. 2010; Zhang et al. 2017; Mittasch et al. 2020), and *let-92* (Enos et al. 2018). We here show that among nematodes these genes are indeed unique among the *Caenorhabditis* genus. Other important factors involved in the special spindle organization might be included in this set of *Caenorhabditis*-specific genes.

Using the amino acid sequences of more than 1500 one-to-one orthologous ORFs with peptide evidence across the 12 species, we generated a phylogeny consolidating already established topologies. Different algorithms of phylogeny reconstruction can vary in their output; thus, we applied three different methods and selected the gene tree that best represents the underlying alignment. These very solid orthology groups represent a highly useful resource for universal nematode analyses given that our study showed that the frequently used BUSCO set of single-copy orthologs does not really represent the common nematode proteome, as reflected in the absence of many of these proteins in *P. pacificus* and *P. redivivus*. Our ProteinOrtho universal orthology groups comprise genes that were found in all species and are highly overlapping with the established nematode BUSCO set, albeit some of them may not be single-copy genes.

Our systematic approach facilitated extensive positive selection analysis of group-specific orthologs able to identify events of evolution that suggest interesting adaptive mechanisms. Very

high frequencies of positive selection were detected for *C. japonica*. The functional enrichments of the positively selected genes are coherent with the special phoretic lifestyle of this nematode that stands out from the other mostly free-living *Caenorhabditis* species. *C. inopinata* shows by far the lowest number of positively selected protein-coding genes. This contrasts with the results of the sister species *C. elegans*, for which we observed positively selected genes to be enriched with muscle-related functions. As these two species are very closely related, this discrepancy is striking. Although *C. inopinata* has only recently been isolated from its natural habitat (Kanzaki et al. 2018), *C. elegans* has been propagated under laboratory conditions for >50 yr now. Previous studies have shown that transferring animals from their natural environments to the laboratory causes strong selective pressures that ultimately can modify the organism genetically and phenotypically (Sterken et al. 2015). Living conditions in the laboratory such as temperature, light, humidity, and oxygen concentration are kept nearly constant; breeding regimes are strictly enforced; and food is unlimited and uniform. In agreement, the phenotype of the laboratory N2 strain of *C. elegans* was shown to be distinct from wild strains in various ways, including aggregation behavior, maturation time, fecundity, body size, and many other traits (De Bono and Bargmann 1998; Kammenga et al. 2007; Weber et al. 2010; Bendesky et al. 2012; Duveau and Félix 2012; Volkers et al. 2013; Andersen et al. 2014; Snoek et al. 2014). When placed in open, liquid-filled, microfluidic chambers containing a square array of posts that mimic complex and structured environments such as soil, *C. elegans* was capable of a novel mode of locomotion, which combines the fast gait of swimming with the more efficient movements of crawling (Park et al. 2008). This mode of locomotion was shown to be very different from the one observed on the smooth surface of agar plates. Also, Gomez-Marin et al. (2016) showed that wild isolates of *C. elegans* show more ordered locomotion than the laboratory reference strain N2. The observed enrichment of muscle-related functions in the *C. elegans* set of positively selected genes might reflect adaptation to distinct requirements for the locomotion on two-dimensional agar plates, as opposed to three-dimensional movement in soil or on rotting fruit (Félix and Braendle 2010). We further observed a widespread adaptive evolution of ribosomal proteins in seven out of the 12 species. Signals of positive selection in individual ribosomal proteins have been previously detected in different organisms (Yednock and Neigel 2014). We here show in a systematic manner that adaptation might in many cases happen at fundamental gene regulatory levels rather than in very specific functional subnetworks. The investigation of such potent evolutionary alterations is of great interest and can be mined in our data (Supplemental Table S10) but will require more experimental validation in the future. Taken together, the results of our study provide annotation improvements and novel evidence for protein-coding genes in diverse nematodes and illustrate our data set to be a valuable genetic resource to facilitate interpretations of biological phenomena through deep phylogenetic comparisons between species that have more recently diverged.

Methods

Nematode culture

The 12 nematode strains (Supplemental Table S1) used in this study were provided by the Caenorhabditis Genetics Center (CGC). Strains were all cultured under the same conditions on nematode growth medium (NGM) plates seeded with *Escherichia coli* OP50 bacteria (Brenner 1974) at 20°C. Nematode cultures were grown until worms of all stages (embryos, larvae, and gravid adults) were visible before bacterial food was exhausted and then processed for RNA and protein extraction as described below.

RNA preparations and RNA sequencing

Mixed worm populations were collected from plates by washing them off the plates with M9 medium, followed by four rounds of spinning and washing. Worm pellets were fast-frozen in 50–100 µL of water in liquid nitrogen and stored at −80°C. For RNA isolation, 500 µL TRIzol LS was added to the frozen pellet and the worms lysed with six freeze–thaw cycles (~30 sec in liquid nitrogen and 2 min in a 37°C water bath; after each cycle, samples were vortexed for 30 sec). Samples were spun down at max speed for 2 min to pellet debris and corpses. The supernatant was transferred to a fresh tube. Then 100% ethanol in a 1:1 ratio was added, mixed well, and pipetted into a Direct-zol RNA miniprep kit (Zymo Research) column. Samples were processed according to the manufacturer's instructions, including the in-column DNase digestion for 30 min. Total RNA was resuspended in 30 µL of RNase-free water. RNA integrity was tested by agarose gel electrophoresis and Bioanalyzer (RNA Nano Assay) and amount-quantified using the Qubit RNA HS assay kit in a Qubit 2.0 fluorometer (Thermo Fisher Scientific). NGS library prep was performed with Illumina's TruSeq stranded mRNA LT sample prep kit following Illumina's standard protocol (part 15031047 rev. E) using one-fourth of the reagents. Libraries were prepared with a starting amount of 250 ng and amplified in 10 PCR cycles. Libraries were profiled using a high sensitivity DNA kit on a 2100 Bioanalyzer (Agilent Technologies) and quantified using the Qubit dsDNA HS assay kit in a Qubit 2.0 fluorometer (Thermo Fisher Scientific). All 12 samples were pooled together in equimolar ratio and sequenced on a NextSeq 500 high output flowcell, PE for 2 × 79 cycles plus seven cycles for the index read. The resulting number of sequenced reads per sample is summarized in Supplemental Table S1.

Protein extraction

Mixed worm populations were collected from plates by washing them off the plates with M9 medium, followed by four rounds of washing. Pellets were fast-frozen in 100 µL water with liquid nitrogen and stored at −80°C. On the day of the protein isolation, samples were thawed and 2× lysis buffer (50 mM Tris-HCl, 300 mM NaCl, 3 mM $MgCl_2$, 2mM DTT, 0.2% Triton X-100, protease inhibitor [cOmplete tablets, mini easypack, Roche]) was added in a 1:1 ratio. Samples were sonicated using a bioruptor plus (Diagenode; 10 cycles 30 sec on and 30 sec off, max intensity). After sonication, the samples were centrifuged at 21,000*g* for 10 min to pellet cell debris. The supernatant was transferred to a fresh reaction tube, and protein concentration of the extract was determined by Bradford (Bio-Rad).

In-gel digestion

In-gel digestion for MS was performed as previously described (Shevchenko et al. 2006). Seventy-five micrograms of each sample was run on a 4%–12% bis-tris gel (Thermo Fisher Scientific) for 40 min at 180 V in 1 × MOPS buffer (Thermo Fisher Scientific). After running, the gel was placed on a clean glass plate, and each sample was sliced into eight pieces with a clean scalpel; each piece was minced and transferred to a 1.5-mL reaction tube. The gel pieces were destained in 50% EtOH/50% ammonium bicarbonate (pH 8.0) buffer at 37°C in a thermoshaker at 1400 rpm until fully destained or slightly blue. After destaining, the gel pieces were

incubated in 100% acetonitrile for 10 min at 25°C, shaking at 1400 rpm until fully dehydrated. The leftover solution was evaporated using a concentrator plus (Eppendorf, settings V-AQ) for 5 min. For reduction, the gel pieces were incubated in 10 mM DTT/50 mM ammonium bicarbonate buffer (pH 8.0) for 60 min at 56°C. Afterward, the gel pieces were incubated with 50 mM iodoacteamide/50 mM ammonium bicarbonate buffer for 45 min at room temperature in the dark. After reduction and alkylation, the gel pieces were washed with 50 mM ammonium bicarbonate buffer (pH 8.0) for 20 min at 25°C, shaking at 1400 rpm. Following the washing step, the gel pieces were again dehydrated in acetonitrile and dried. To digest the proteins, the dried gel pieces were rehydrated with 50 mM ammonium bicarbonate buffer (pH 8) containing 1 µg MS-grade trypsin (Sigma-Aldrich) and incubated overnight at 37°C. The supernatant of trypsin solution was recovered and saved in a fresh reaction tube. Tryptic peptides were extracted from the gel pieces by incubation with 30% acetonitrile twice for 15 min at 25°C, shaking at 1400 rpm. The supernatant was recovered each time and combined with the previously recovered fractions. Finally, the gel pieces were dehydrated by incubation in acetonitrile until fully dry. The acetonitrile was recovered and combined with the previously collected supernatants. The sample solution containing the tryptic peptides was reduced to 10% original volume in a concentrator plus (Eppendorf, settings V-AQ).

Stage tip purification

Stage tip purification was performed as previously described (Rappsilber et al. 2007). Desalting tips were prepared by stacking two layers of Empore C18 material (3M) in a 200-µL pipette tip. After activation of the tips with pure methanol, spinning at 500*g*, they were washed two times with 80% acetonitrile/0.1% formic acid and then with 0.1% formic acid for 5 min at 500*g*. The tryptic peptide samples were applied and centrifuged at 500*g*. After one more wash with 0.1% formic acid, the peptides were eluted into a 24-well plate (Thermo Fisher Scientific) with 80% acetonitrile/0.1% formic acid by centrifugation at 500*g* for 3 min. To evaporate the acetonitrile, the samples were concentrated in a concentrator plus (Eppendorf, setting V-AQ) for 10 min and finally filled up to 14 µL with 50 mM ammonium bicarbonate (pH 8)/0.1% formic acid. Half the volume of the samples was measured on the MS, whereas the other half was stored at −20°C as backup.

Mass spectrometry measurements

Peptides were analyzed by nanoflow liquid chromatography either on an EASY-nLC 1000 system (Thermo Scientific) coupled to a Q Exactive plus mass spectrometer (Thermo Scientific) or an EASY-nLC 1200 system (Thermo Scientific) coupled to an Exploris 480 (Thermo Scientific). Peptides were separated on a C18-reversed-phase column (20-cm or 60-cm length, 75-µm diameter) packed in-house with Reprosil aq1.9 (Dr. Maisch GmbH), directly mounted on the electrospray ion source of the mass spectrometer. For both HPLC systems, peptides were eluted from the column in an optimized 103-min (Exploris) and 208-min (QEP) gradient from 2% to 40% with a mixture of 80% acetonitrile/0.1% formic acid at a flow rate of 225–250 nL/min. The QEP was operated in positive ion mode with a data-dependent acquisition strategy of one MS full scan (scan range 300–1650 m/z; 70,000 resolution; AGC target 3e6; max IT 20 msec) and up to 10 MS/MS scans (17,500 resolution; AGC target 1e5, max IT 120 msec; isolation window 1.8 m/z) with peptide match preferred using HCD fragmentation. The Exploris was operated in positive ion mode with a data-dependent acquisition strategy of one MS full scan (scan range 300–1650 m/z; 60,000 resolution; normalized AGC target 300%; max IT 28 msec) and up to 20 MS/MS scans (15,000 resolution; AGC target 100%, max IT 28 msec; isolation window 1.4 m/z) with peptide match preferred using HCD fragmentation.

Transcriptome assembly

The Illumina 79 bases paired-end RNA-seq data sets were used to assemble the transcriptome. First, erroneous *k*-mers were removed using Rcorrector (Song and Florea 2015) and the specialized scripts from TranscriptomeAssemblyTools (FilterUncorrectablePEfastq.py). Second, adapter sequences were trimmed using Trim Galore! (a wrapper around cutadapt [Martin 2011] and FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/]), and reads were filtered to include only pairs of minimum length of 36 nt each. These clean-up steps removed only 1% of the paired reads. The remaining corrected reads were cleaned from reads that might stem from the food source *E. coli* by mapping the reads to the *E. coli* genome (downloaded from the NCBI Assembly database [https://www.ncbi.nlm.nih.gov/assembly] under GCF_000005845.2_ASM584v2_genomic.fna.gz) using STAR (version 2.5.4b) (Dobin et al. 2013), and only unmapped reads were used for the next steps. For the GG assembly, corrected raw RNA-seq data were mapped to the respective genomes (Supplemental Table S1) using STAR (version 2.5.4b) (Dobin et al. 2013). The corrected raw RNA-seq or mapped data were used for GF de novo or GG assembly approach using the Trinity suite (Trinity version 2.4.0) (Grabherr et al. 2011) with the following parameter setting: for GF, --seqType fq --SS_lib_type RF --min_kmer_cov 1; and for GG, Trinity --genome_guided_bam --genome_guided_max_intron 3500 --genome_guided_min_coverage 2. The maximum intron size is needed as a parameter for STAR alignment and Trinity assembly and was determined based on previous work (Wu et al. 2013). The resulting Trinity FASTA files were then further processed with TransDecoder (version 5.4.0) (Bryant et al. 2017; http://transdecoder.github.io) to predict potential protein-coding transcripts using a length threshold of 20 amino acids. The resulting peptide FASTA files were used as search space in subsequent steps for mass spectrometry data analysis.

Assembly quality assessment

The quality of the assembled transcriptome was assessed using several different state-of-the-art approaches. These included general metrics of number of assembled transcripts, mean, median, and Ex90N50 transcript lengths. The alignment rate of the raw reads to the assembly was calculated using Bowtie 2 (version 2.3.4.3) (Langmead and Salzberg 2012) and dedicated scripts provided by Trinity (version 2.4.0) (Grabherr et al. 2011). BUSCO (version 5.0.0) (Simão et al. 2015) was used to assess transcriptome completeness in both assemblies (GF and GG). The testing model was "protein," and we used a set of 3131 BUSCO groups of universal single-copy orthologs of the "nematoda_odb10 database." TransRate scores and additional quality metrics were established using TransRate (version 1.0.3) (Smith-Unna et al. 2016). Coherence with current annotations was measured using a combination of BLASTP (BLAST+ version 2.8.1) (Camacho et al. 2009) and Trinity tools (version 2.4.0) (Grabherr et al. 2011). For RNA-seq coverage validations, the combined cleaned RNA-seq data were mapped to the respective genome assembly using STAR (version 2.5.4b) (Dobin et al. 2013). Assembly efficiency as depicted in Supplemental Figure S9A was calculated by dividing the number of assembled contigs that pass the machine-learning-predicted completeness of 80% by the number of sequenced reads used for the assembly. WormBase genome assemblies and annotations were

assayed for genome content (relevant for Supplemental Fig. S1) using the respective annotation GFF3 files and the agat_sp_statistics.pl tool from the AGAT GTF/GFF analysis toolkit (https://github.com/NBISweden/AGAT).

Annotation of identified transcripts

Functional and domain annotations were produced using Trinotate (version 3.1.1) (Bryant et al. 2017; https://github.com/Trinotate/Trinotate/wiki) combining the following applications: HMMER (version 3.2.1) (Eddy 2011) to identify protein domains, signalP (version 5.0) (Almagro Armenteros et al. 2019) to predict signal peptides, TMHMM (version 2.0c) (Krogh et al. 2001) to predict transmembrane regions, RNAMMER (version 1.2) (Lagesen et al. 2007) to identify rRNA transcripts in addition to infer Gene Ontology, and KEGG terms from orthologs established by BLAST+ (version 2.8.1) (Camacho et al. 2009) with a Swiss-Prot database of all major model species. Further, localization predictions from protein sequences of the assembly were calculated using DeepLoc (version 1.0) (Almagro Armenteros et al. 2017).

Genome annotation sources

Genome sequence, proteome, and gene annotations for nine nematode species (*C. elegans*, *C. briggsae*, *C. brenneri*, *C. japonica*, *C. remanei*, *P. pacificus*, *O. tipulae*, *P. redivivus*, and *C. inopinata*) were downloaded from WormBase version WS273 (Harris et al. 2010).

Protein identification and label-free quantification

MaxQuant (version 1.6.5.0) (Cox and Mann 2008) was used for raw file peak extraction and protein identification against the respective Trinity GF, Trinity GG, or WormBase protein FASTA files. The proteome of *E. coli* (strain K12) from UniProt (Proteome ID UP000000625 (version August 21, 2019) was included for filtering *E. coli* contaminants. Protein quantification was performed with MaxQuant using the label-free quantification (LFQ) algorithm (Cox et al. 2014). The following parameters were applied: trypsin as cleaving enzyme; minimum peptide length of seven amino acids; maximal two missed cleavages; carbamidomethylation of cysteine as a fixed modification; N-terminal protein acetylation; and oxidation of methionine as variable modifications. Further settings were "label-free quantification" with "FastLFQ" disabled, "match between runs" with a time window of 0.7 min for matching and 20 min for alignment; peptide and protein false-discovery rates (FDR) were set to 0.01; and common contaminants (provided via standard MaxQuant contaminant list) were excluded. Detailed settings are available in the respective parameter files uploaded to ProteomeXchange. MaxQuant LFQ data were further processed using in-house-developed tools based on R (version 3.5.3) (R Core Team 2022). This included filtering out marked contaminants, *E. coli*–specific proteins, reverse entries, and proteins only identified by site. Protein groups with no unique and fewer than two peptides were removed. Before imputation of missing LFQ values with a β-distribution ranging from 0.1 to 0.2 percentile within each sample, the values were $\log_2$-transformed.

Machine learning for transcript completeness prediction

We reckoned that transcript completeness could most probably be predicted by combining different measures of how well the underlying reads support the assembled transcript. We hence implemented random forest (RF) of the "caret" R package (Kuhn 2008) with default parameters using *C. elegans* assembly quality measurements from TransRate software (Smith-Unna et al. 2016) and transcript features provided by TransDecoder (https://github.com/TransDecoder/TransDecoder) as features (for detailed information, see Supplemental Table S4) and BLASTP percentage hit length representing transcript completeness as the target variable to train regression models. At each of the 500 iterations of the cross-validation, 75% of the input values was used to build the subtraining set, and the remaining 25% (subtesting set) was tested. Using assembly mode–specific models for GF and GG assemblies, we predicted transcript completeness of ORFs in all species and both modes using the respective TransRate assembly measures and transcript features. To assess applicability also in other species, we assembled publicly available RNA sequencing data of other well-studied model organisms, including the nematode *C. briggsae*, the fruit fly *D. melanogaster*, the green land plant *A. thaliana*, and the human H1-hESC cell line (Supplemental Table S5), using the same workflows and applied the two RF models.

Enrichment analysis (ontology and pathways)

All relevant gene lists were interrogated for functional enrichments using functional annotation from various databases such as KEGG pathways (Ogata et al. 1999), Gene Ontology (Ashburner et al. 2000), Pfam (Sonnhammer et al. 1998), SMART (Schultz et al. 1998), and knockout phenotype. Fisher's exact test was applied to the respective gene lists and the background set of genes that varied depending on which data set was analyzed. For the *Caenorhabditis*-specific genes, the background consisted of all WormBase *C. elegans* genes that were included in any orthologous group. For species-specific positive selection genes, we used all genes that were interrogated for positive selection as background list of genes.

Enrichment analysis with STRINGdb

To enable functional interpretation of certain lists of genes we interrogated them using the online tool STRINGdb version 11.5 (Szklarczyk et al. 2021), which enables the identification of protein–protein networks and functional enrichment analysis. We excluded association data of "textmining" and "co-occurrence" sources. In the network display, we chose to hide disconnected nodes. All other settings were kept as default. The resulting gene network was clustered with the built-in MCL clustering with an inflation parameter of 3.1.

Analysis of expression pattern of new *C. elegans* genes

Raw reads of two *C. elegans* NGS time course studies (Boeck et al. 2016; Levin et al. 2016) were downloaded from the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra). Reads were mapped to the *C. elegans* reference genome c_elegans.PRJNA13758.WS273.genomic.fa together with the accompanying gene models from WormBase version WS273 and the genomic features of the two suggested new *C. elegans* genes using STAR version 2.5.1b (Dobin et al. 2013) and allowing up to two mismatches. Only uniquely mapped reads were used to quantify expression of genes, using featureCounts v1.4.6 p2 (Liao et al. 2013) with default parameters and the same gene model used for mapping. "Fragments per million" values of the individual ORFs were calculated using the fpm function from the DESeq2 R package (Love et al. 2014).

Protein orthology search

We used ProteinOrtho v6.06 (Lechner et al. 2011) to establish orthologous groups across the 12 species. We used ProteinOrtho with default parameters in two data sets: (1) the whole transcriptome including all transcripts with RF completeness prediction of >80% and (2) ORFs with peptide evidence and RF completeness

prediction of >80%. This resulted in 23,090 orthologous groups for transcriptome and 14,261 for ORFs with peptide evidence. Furthermore, we included the protein information from *C. elegans* WormBase in the analysis for annotation to allow for functional annotation from various databases such as KEGG pathways (Ogata et al. 1999), Gene Ontology (Ashburner et al. 2000), Pfam (Sonnhammer et al. 1998), SMART (Schultz et al. 1998), knockout phenotype, and subsequent enrichment analyses.

Determination of positive selection

From the ProteinOrtho-established orthology groups, we extracted only 1:1 orthologous gene clusters that included at least three species. The respective amino acid and CDS sequences were retrieved from the TransDecoder output files and aligned using PRANK (version 170427) (Löytynoja and Goldman 2008), which has been used in other evolutionary analysis (Fletcher and Yang 2010). "Reverse translation" to obtain the accurate codon alignment was performed using PAL2NAL.v14 with "removing gaps," "in-frame stop codons," and "mismatched codons" settings (Suyama et al. 2006). Evolutionary rates were estimated using the PAML CODEML program (Yang 2007). The ProteinOrtho groups were fitted to six different models (lineage-specific models and site-specific substitution models) for detecting codons under positive or purifying selection or drift. The rate of protein evolution was estimated with model M0 (one ratio), which assumes that all amino acid sites have a single value of ω. Positively selected sites were identified based on two pairs of models: nearly neutral models (M1a and M7) and positive selection models (M8 and M2a). M1a (nearly neutral) assumes two classes of sites ($\omega = 1$, $0 < \omega < 1$); M2a (positive selection) assumes three site classes ($\omega = 1$, $0 < \omega < 1$, and $\omega > 1$); and M3 (discrete) assumes three discrete distributions of three site classes, with different ω values estimated from the data. M7 (β) assumes a β-distribution of class sites for 10 different ω ratios in the interval (0, 1) that does not allow for selection ($0 < \omega < 1$), and M8 (β and ω; continuous) adds an extra class of sites with positive selection ($\omega > 1$) to the β (M7) model (Nielsen and Yang 1998). For each included ProteinOrtho group, we computed the likelihood ratio tests (LRTs) pairing models M1 with M2 and M7 with M8 and selected any group that had a log-likelihood score 2ΔLnL difference of at least two between the two models for further analysis. We subsequently retrieved the *P*-value by comparing each 2ΔLnL against the χ^2 distribution using the respective degrees of freedom (df) of each model pair. *P*-values were corrected for multiple testing using the Benjamini–Hochberg method. A ProteinOrtho group was considered to be undergoing site-specific diversifying selection if the LRT result was significant (FDR < 0.05). We determined the model pair with highest likelihood to identify orthology groups that show evidence for positive selection and found the M7 versus M8 model comparison to consistently provide highest significance compared with the M1 versus M2 comparisons. This trend has already been described by others (Anisimova et al. 2001; Wong et al. 2004). Subsequently, the posterior probabilities of each codon belonging to the site class of positive selection ($\omega > 1$) were estimated with the Bayes empirical Bayes (BEB) method (Yang et al. 2005). To detect branch-specific positive selection for each ProteinOrtho group with site-specific positive selection signals, we applied the LRT-based branch-specific and branch-site-specific models across the different species in the phylogenetic tree, dividing the tree into all possible combinations of one of the terminal branches as the foreground branch and the remaining as background branches. This results in the same number of calculations as the number of orthologs in the inspected ProteinOrtho group (minimum, three; maximum, 12). The significance of the LRTs was calculated assuming a constant ω across all sites and branches of the respective phylogeny using the M0 model (Nielsen and Yang 1998). *P*-values were corrected for multiple testing using the Benjamini–Hochberg method, and branch-site-specific positive selection signals with FDR < 0.1 were reported as significant and further analyzed. Terminal branch average d_N/d_S values as depicted in Supplemental Figure S9C were calculated from the terminal branch d_N/d_S values provided by CODEML, including all one-to-one orthologs across all species.

Phylogenetic relation analyses

Multiple sequence alignments of one-to-one orthology groups as established for the positive selection analysis were used to reconstruct individual gene trees by performing maximum likelihood (ML) analyses with the phylogenetic analysis tools RAxML (version 8.2.12) and FastTree (version 2.1.10).

The following commands were used to run these programs:

RAxML (and RAxML-Limited):
raxmlHPC -f a -m GTRGAMMA -p 12345 -x 12345 -# 100 -s <input_alignment> -n <output_tree_1>
raxmlHPC -f a -m GTRGAMMA -p 23456 -x 23456 -# 100 -s <input_alignment >-n <output_tree_2>
FastTree:
FastTree -nt -gtr -nosupport -log <log file> <input_alignment> > <output_tree_3>

Using these commands, we reconstructed three individual gene trees for each one-to-one ortholog group, selected the best based on the maximum likelihood scores of the individual trees, and finally summarized all individual gene trees into an unrooted phylogenetic species tree using ASTRAL (version 5.7.8) (Mirarab et al. 2014).

Data access

All raw RNA-seq data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA843607. The mass spectrometry proteomics data generated in this study have been submitted to the ProteomeXchange Consortium (http://www.proteomexchange.org) via the Proteomics Identifications Database (PRIDE) (Perez-Riverol et al. 2022) partner repository with the data set identifier PXD034107. All Trinity assemblies, TransDecoder CDS and peptide files, and the ProteinOrtho tables are provided in Supplemental Material.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Jasmin Cartano and Franziska Roth for excellent technical assistance. Transcriptome samples were processed and measured by the Genomics Core Facility at IMB. This project was funded by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) GRK2526/1–Projectnr. 407023052.

References

Ahmed M, Roberts NG, Adediran F, Smythe AB, Kocot KM, Holovachov O. 2021. Phylogenomic analysis of the phylum Nematoda: conflicts and congruences with morphology, 18S rRNA and mitogenomes. *Front Ecol Evol* **9.** doi:10.3389/fevo.2021.769565

Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. 2017. DeepLoc: prediction of protein subcellular localization

using deep learning. *Bioinformatics* **33:** 3387–3395. doi:10.1093/BIOINFORMATICS/BTX431
Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* **37:** 420–423. doi:10.1038/S41587-019-0036-Z
Andersen EC, Bloom JS, Gerke JP, Kruglyak L. 2014. A variant in the neuropeptide receptor *npr-1* is a major determinant of *Caenorhabditis elegans* growth and physiology. *PLoS Genet* **10:** e1004156. doi:10.1371/JOURNAL.PGEN.1004156
Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* **18:** 1585–1592. doi:10.1093/OXFORDJOURNALS.MOLBEV.A003945
Antoshechkin I, Sternberg PW. 2007. The versatile worm: genetic and genomic resources for *Caenorhabditis elegans* research. *Nat Rev Genet* **8:** 518–532. doi:10.1038/NRG2105
Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS. 2000. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25–29. doi:10.1038/75556
Bayliss R, Sardon T, Vernos I, Conti E. 2003. Structural basis of Aurora-A activation by TPX2 at the mitotic spindle. *Mol Cell* **12:** 851–862. doi:10.1016/S1097-2765(03)00392-7
Bendesky A, Pitts J, Rockman MV, Chen WC, Tan MW, Kruglyak L, Bargmann CI. 2012. Long-range regulatory polymorphisms affecting a GABA receptor constitute a quantitative trait locus (QTL) for social behavior in *Caenorhabditis elegans*. *PLoS Genet* **8:** e1003157. doi:10.1371/JOURNAL.PGEN.1003157
Blaxter M. 2016. Imagining Sisyphus happy: DNA barcoding and the unnamed majority. *Philos Trans R Soc Lond B Biol Sci* **371:** 20150329. doi:10.1098/RSTB.2015.0329
Boeck ME, Huynh C, Gevirtzman L, Thompson OA, Wang G, Kasper DM, Reinke V, Hillier LW, Waterston RH. 2016. The time-resolved transcriptome of *C. elegans*. *Genome Res* **26:** 1441–1450. doi:10.1101/GR.202663.115
Bongers T, Bongers M. 1998. Functional diversity of nematodes. *Appl Soil Ecol* **10:** 239–251. doi:10.1016/S0929-1393(98)00123-1
Brenner S. 1973. The genetics of behaviour. *Br Med Bull* **29:** 269–271. doi:10.1093/OXFORDJOURNALS.BMB.A071019
Brenner S. 1974. The genetics of *Caenorhabditis elegans*. *Genetics* **77:** 71–94. doi:10.1093/GENETICS/77.1.71
Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, Lee TJ, Leigh ND, Kuo TH, Davis FG, et al. 2017. A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. *Cell Rep* **18:** 762–776. doi:10.1016/J.CELREP.2016.12.063
Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10:** 421. doi:10.1186/1471-2105-10-421
Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP. 2008. Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc Natl Acad Sci* **105:** 21034–21038. doi:10.1073/PNAS.0811066106
Chapman B, Bellgard M. 2017. Plant proteogenomics: improvements to the grapevine genome annotation. *Proteomics* **17:** 1700197. doi:10.1002/PMIC.201700197
Choe A, Von Reuss SH, Kogan D, Gasser RB, Platzer EG, Schroeder FC, Sternberg PW. 2012. Ascaroside signaling is widely conserved among nematodes. *Curr Biol* **22:** 772–780. doi:10.1016/J.CUB.2012.03.024
Conduit PT, Wainman A, Raff JW. 2015. Centrosome function and assembly in animal cells. *Nat Rev Mol Cell Biol* **16:** 611–624. doi:10.1038/NRM4062
Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26:** 1367–1372. doi:10.1038/NBT.1511
Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. 2014. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13:** 2513–2526. doi:10.1074/MCP.M113.031591
Cutter AD. 2008. Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. *Mol Biol Evol* **25:** 778–786. doi:10.1093/MOLBEV/MSN024
Dalley BK, Golomb M. 1992. Gene expression in the *Caenorhabditis elegans* dauer larva: developmental regulation of Hsp90 and other genes. *Dev Biol* **151:** 80–90. doi:10.1016/0012-1606(92)90215-3
Davidson EH, Erwin DH. 2006. Gene regulatory networks and the evolution of animal body plans. *Science* **311:** 796–800. doi:10.1126/SCIENCE.1113832
De Bono M, Bargmann CI. 1998. Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in *C. elegans*. *Cell* **94:** 679–689. doi:10.1016/S0092-8674(00)81609-8
Delattre M, Goehring NW. 2021. The first steps in the life of a worm: themes and variations in asymmetric division in *C. elegans* and other nematodes. *Curr Top Dev Biol* **144:** 269–308. doi:10.1016/BS.CTDB.2020.12.006
Desgagné-Penix I, Khan MF, Schriemer DC, Cram D, Nowak J, Facchini PJ. 2010. Integration of deep transcriptome and proteome analyses reveals the components of alkaloid metabolism in opium poppy cell cultures. *BMC Plant Biol* **10:** 252. doi:10.1186/1471-2229-10-252
Ding N, Zhang B, Ying W, Song J, Feng L, Zhang K, Li H, Xu J, Xiao T, Cheng S. 2020. A time-resolved proteotranscriptomics atlas of the human placenta reveals pan-cancer immunomodulators. *Signal Transduct Target Ther* **5:** 110. doi:10.1038/S41392-020-00224-5
Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/BIOINFORMATICS/BTS635
Duveau F, Félix MA. 2012. Role of pleiotropy in the evolution of a cryptic developmental variation in *Caenorhabditis elegans*. *PLoS Biol* **10:** e1001230. doi:10.1371/JOURNAL.PBIO.1001230
Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol* **7:** e1002195. doi:10.1371/JOURNAL.PCBI.1002195
Enos SJ, Dressler M, Gomes BF, Hyman AA, Woodruff JB. 2018. Phosphatase PP2A and microtubule-mediated pulling forces disassemble centrosomes during mitotic exit. *Biol Open* **7:** bio029777. doi:10.1242/BIO.029777/VIDEO-6
Evans VC, Barker G, Heesom KJ, Fan J, Bessant C, Matthews DA. 2012. *De novo* derivation of proteomes from transcriptomes for transcript and protein identification. *Nat Methods* **9:** 1207–1211. doi:10.1038/NMETH.2227
Félix MA, Braendle C. 2010. The natural history of *Caenorhabditis elegans*. *Curr Biol* **20:** R965–R969. doi:10.1016/J.CUB.2010.09.050
Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol* **27:** 2257–2267. doi:10.1093/MOLBEV/MSQ115
Gomez-Marin A, Stephens GJ, Brown AE. 2016. Hierarchical compression of *Caenorhabditis elegans* locomotion reveals phenotypic differences in the organization of behaviour. *Journal of The Royal Society Interface* **13:** 20160466. doi:10.1098/rsif.2016.0466
Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29:** 644–652. doi:10.1038/NBT.1883
Gupta BP, Johnsen R, Chen N. 2007. Genomics and biology of the nematode *Caenorhabditis briggsae*. *WormBook* **May 3:** 1–16. doi:10.1895/wormbook.1.136.1
Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8:** 1494–1512. doi:10.1038/NPROT.2013.084
Hamill DR, Severson AF, Carter JC, Bowerman B. 2002. Centrosome maturation and mitotic spindle assembly in *C. elegans* require SPD-5, a protein with multiple coiled-coil domains. *Dev Cell* **3:** 673–684. doi:10.1016/S1534-5807(02)00327-1
Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol* **30:** 1987–1997. doi:10.1093/MOLBEV/MST100
Hao Y, Zhang L, Niu Y, Cai T, Luo J, He S, Zhang B, Zhang D, Qin Y, Yang F, et al. 2018. SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief Bioinform* **19:** 636–643. doi:10.1093/BIB/BBX005
Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, De la Cruz N, Davis P, Duesbury M, Fang R, et al. 2010. WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res* **38:** D463–D467. doi:10.1093/NAR/GKP952
Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Cho J, Davis P, Gao S, Grove CA, Kishore R, et al. 2020. WormBase: a modern Model Organism Information Resource. *Nucleic Acids Res* **48:** D762–D767. doi:10.1093/NAR/GKZ920
Henen MA, Myers W, Schmitt LR, Wade KJ, Born A, Nichols PJ, Vögeli B. 2021. The disordered spindly C-terminus interacts with RZZ subunits ROD-1 and ZWL-1 in the kinetochore through the same sites in *C. elegans*. *J Mol Biol* **433:** 166812. doi:10.1016/J.JMB.2021.166812
Hillier LDW, Coulson A, Murray JI, Bao Z, Sulston JE, Waterston RH. 2005. Genomics in *C. elegans*: so many genes, such a little worm. *Genome Res* **15:** 1651–1660. doi:10.1101/GR.3729105
Hodda M, Peters L, Traunspurger W. 2009. Nematode diversity in terrestrial, freshwater aquatic and marine systems. In *Nematodes as environmental indicators* (ed. Wilson MJ, Kakouli-Duarte T), pp. 45–93. CABI, Oxfordshire, UK. doi:10.1079/9781845933852.0045
Horvitz HR. 2003. Worms, life, and death (Nobel lecture). *Chembiochem* **4:** 697–711. doi:10.1002/CBIC.200300614

Howe K, Davis P, Paulini M, Tuli MA, Williams G, Yook K, Durbin R, Kersey P, Sternberg PW. 2012. WormBase: annotating many nematode genomes. *Worm* **1:** 15–21. doi:10.4161/WORM.19574

Jaffe JD, Berg HC, Church GM. 2004. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4:** 59–77. doi:10.1002/PMIC.200300511

Kaletta T, Hengartner MO. 2006. Finding function in novel targets: *C. elegans* as a model organism. *Nat Rev Drug Discov* **5:** 387–399. doi:10.1038/NRD2031

Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, Ohler U, Bergman CM, Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468:** 811–816. doi:10.1038/NATURE09634

Kammenga JE, Doroszuk A, Riksen JAG, Hazendonk E, Spiridon L, Petrescu AJ, Tijsterman M, Plasterk RHA, Bakker J. 2007. A *Caenorhabditis elegans* wild type defies the temperature-size rule owing to a single nucleotide polymorphism in *tra-3*. *PLoS Genet* **3:** e34. doi:10.1371/JOURNAL.PGEN.0030034

Kanzaki N, Tsai IJ, Tanaka R, Hunt VL, Liu D, Tsuyama K, Maeda Y, Namai S, Kumagai R, Tracey A, et al. 2018. Biology and genome of a newly discovered sibling species of *Caenorhabditis elegans*. *Nat Commun* **9:** 3216. doi:10.1038/S41467-018-05712-5

Kiontke K, Fitch DHA. 2013. Nematodes. *Curr Biol* **23:** R862–R864. doi:10.1016/J.CUB.2013.08.009

Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305:** 567–580. doi:10.1006/JMBI.2000.4315

Kuhn M. 2008. Building predictive models in R using the caret package. *J Stat Softw* **28:** 1–26. doi:10.18637/jss.v028.i05

Kumar D, Yadav AK, Jia X, Mulvenna J, Dash D. 2016. Integrated transcriptomic-proteomic analysis using a proteogenomic workflow refines rat genome annotation. *Mol Cell Proteomics* **15:** 329–339. doi:10.1074/MCP.M114.047126

Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35:** 3100–3108. doi:10.1093/NAR/GKM160

Lang X, Li N, Li L, Zhang S. 2019. Integrated metabolome and transcriptome analysis uncovers the role of anthocyanin metabolism in *Michelia maudiae*. *Int J Genomics* **2019:** 4393905. doi:10.1155/2019/4393905

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359. doi:10.1038/NMETH.1923

Larsen PL. 1993. Aging and resistance to oxidative damage in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A* **90:** 8905–8909. doi:10.1073/PNAS.90.19.8905

Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**. doi:10.1186/1471-2105-12-124

Leung MCK, Williams PL, Benedetto A, Au C, Helmcke KJ, Aschner M, Meyer JN. 2008. *Caenorhabditis elegans*: an emerging model in biomedical and environmental toxicology. *Toxicol Sci* **106:** 5–28. doi:10.1093/TOXSCI/KFN121

Levin M, Butter F. 2022. Proteotranscriptomics: a facilitator in omics research. *Comput Struct Biotechnol J* **20:** 3667–3675. doi:10.1016/J.CSBJ.2022.07.007

Levin M, Anavy L, Cole AG, Winter E, Mostov N, Khair S, Senderovich N, Kovalev E, Silver DH, Feder M, et al. 2016. The mid-developmental transition and the evolution of animal body plans. *Nature* **531:** 637–641. doi:10.1038/NATURE16994

Levin M, Scheibe M, Butter F. 2020. Proteotranscriptomics assisted gene annotation and spatial proteomics of *Bombyx mori* BmN4 cell line. *BMC Genomics* **21**. doi:10.1186/S12864-020-07088-7

Liao Y, Smyth GK, Shi W. 2013. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* **41:** e108. doi:10.1093/NAR/GKT214

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550. doi:10.1186/S13059-014-0550-8

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320:** 1632–1635. doi:10.1126/SCIENCE.1158395

Ma J, Saghatelian A, Shokhirev MN. 2018. The influence of transcript assembly on the proteogenomics discovery of microproteins. *PLoS One* **13:** e0194518. doi:10.1371/JOURNAL.PONE.0194518

Magescas J, Zonka JC, Feldman JL. 2019. A two-step mechanism for the inactivation of microtubule organizing center function at the centrosome. *eLife* **8:** e47867. doi:10.7554/eLife.47867

Malik A, Gildor T, Sher N, Layous M, Ben-Tabou de-Leon S. 2017. Parallel embryonic transcriptional programs evolve under distinct constraints and may enable morphological conservation amidst adaptation. *Dev Biol* **430:** 202–213. doi:10.1016/J.YDBIO.2017.07.019

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17:** 10–12. doi:10.14806/ej.17.1.200

Melsted P, Hateley S, Joseph IC, Pimentel H, Bray N, Pachter L. 2017. Fusion detection and quantification by pseudoalignment. bioRxiv doi:10.1101/166322

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30:** i541–i548. doi:10.1093/BIOINFORMATICS/BTU462

Mittasch M, Tran VM, Rios MU, Fritsch AW, Enos SJ, Gomes BF, Bond A, Kreysing M, Woodruff JB. 2020. Regulated changes in material properties underlie centrosome disassembly during mitotic exit. *J Cell Biol* **219:** e201912036. doi:10.1083/JCB.201912036

Mohien CU, Colquhoun DR, Mathias DK, Gibbons JG, Armistead JS, Rodriguez MC, Rodriguez MH, Edwards NJ, Hartler J, Thallinger GG, et al. 2013. A bioinformatics approach for integrated transcriptomic and proteomic comparative analyses of model and non-sequenced anopheline vectors of human malaria parasites. *Mol Cell Proteomics* **12:** 120–131. doi:10.1074/MCP.M112.019596

Mukherjee K, Bürglin TR. 2007. Comprehensive analysis of animal TALE homeobox genes: new conserved motifs and cases of accelerated evolution. *J Mol Evol* **65:** 137–153. doi:10.1007/S00239-006-0023-0

Müller T, Boileau E, Talyan S, Kehr D, Varadi K, Busch M, Most P, Krijgsveld J, Dieterich C. 2021. Updated and enhanced pig cardiac transcriptome based on long-read RNA sequencing and proteomics. *J Mol Cell Cardiol* **150:** 23–31. doi:10.1016/J.YJMCC.2020.10.005

Müller-Reichert T, Greenan G, O'Toole E, Srayko M. 2010. The *elegans* of spindle assembly. *Cell Mol Life Sci* **67:** 2195–2213. doi:10.1007/S00018-010-0324-8

Nasa I, Kettenbach AN. 2018. Coordination of protein kinase and phosphoprotein phosphatase activities in mitosis. *Front Cell Dev Biol* **6:** 30. doi:10.3389/FCELL.2018.00030

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148:** 929–936. doi:10.1093/GENETICS/148.3.929

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **27:** 29–34. doi:10.1093/NAR/27.1.29

O'Riordan VB, Burnell AM. 1990. Intermediary metabolism in the dauer larva of the nematode *Caenorhabditis elegans* II: the glyoxylate cycle and fatty-acid oxidation. *Comp Biochem Physiol B* **95:** 125–130. doi:10.1016/0305-0491(90)90258-U

Park S, Hwang H, Nam SW, Martinez F, Austin RH, Ryu WS. 2008. Enhanced *Caenorhabditis elegans* locomotion in a structured microfluidic environment. *PLoS One* **3:** e2550. doi:10.1371/journal.pone.0002550

Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, Kundu DJ, Prakash A, Frericks-Zipper A, Eisenacher M, et al. 2022. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* **50:** D543–D552. doi:10.1093/NAR/GKAB1038

Prasad TSK, Mohanty AK, Kumar M, Sreenivasamurthy SK, Dey G, Nirujogi RS, Pinto SM, Madugundu AK, Patil AH, Advani J, et al. 2017. Integrating transcriptomic and proteomic data for accurate assembly and annotation of genomes. *Genome Res* **27:** 133–144. doi:10.1101/GR.201368.115

Rappsilber J, Mann M, Ishihama Y. 2007. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc* **2:** 1896–1906. doi:10.1038/NPROT.2007.261

R Core Team. 2022. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Rödelsperger C. 2021. The community-curated *Pristionchus pacificus* genome facilitates automated gene annotation improvement in related nematodes. *BMC Genomics* **22:** 216. doi:10.1186/S12864-021-07529-X

Rödelsperger C, Athanasouli M, Lenuzzi M, Theska T, Sun S, Dardiry M, Wighard S, Hu W, Sharma DR, Han Z. 2019. Crowdsourcing and the feasibility of manual gene annotation: a pilot study in the nematode *Pristionchus pacificus*. *Sci Rep* **9:** 18789. doi:10.1038/S41598-019-55359-5

Romiguier J, Figuet E, Galtier N, Douzery EJ, Boussau B, Dutheil JY, Ranwez V. 2012. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. *PLoS One* **7:** e33852. doi:10.1371/journal.pone.0033852

Schlaitz AL, Srayko M, Dammermann A, Quintin S, Wielsch N, MacLeod I, de Robillard Q, Zinke A, Yates JR, Müller-Reichert T, et al. 2007. The *C. elegans* RSA complex localizes protein phosphatase 2A to centrosomes and regulates mitotic spindle assembly. *Cell* **128:** 115–127. doi:10.1016/J.CELL.2006.10.050

Schultz J, Milpetz F, Bork P, Ponting CP. 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci* **95:** 5857–5864. doi:10.1073/PNAS.95.11.5857

Shevchenko A, Tomas H, Havliš J, Olsen JV, Mann M. 2006. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat Protoc* **1:** 2856–2860. doi:10.1038/NPROT.2006.468

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31:** 3210–3212. doi:10.1093/BIOINFORMATICS/BTV351

Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. 2016. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res* **26:** 1134–1144. doi:10.1101/GR.196469.115

Snoek LB, Sterken MG, Volkers RJM, Klatter M, Bosman KJ, Bevers RPJ, Riksen JAG, Smant G, Cossins AR, Kammenga JE. 2014. A rapid and massive gene expression shift marking adolescent transition in *C. elegans*. *Sci Rep* **4:** 3912. doi:10.1038/SREP03912

Song L, Florea L. 2015. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* **4:** 48. doi:10.1186/S13742-015-0089-Y

Sonnhammer ELL, Eddy SR, Birney E, Bateman A, Durbin R. 1998. Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **26:** 320–322. doi:10.1093/NAR/26.1.320

Stenzel L, Mehler J, Schreiner A, Uestuener S, Zuccoli E, Zanin E, Mikeladze-Dvali T. 2021. PCMD-1 bridges the centrioles and the pericentriolar material scaffold in *C. elegans*. *Development* **148:** dev198416. doi:10.1242/DEV.198416

Sterken MG, Snoek LB, Kammenga JE, Andersen EC. 2015. The laboratory domestication of *Caenorhabditis elegans*. *Trends Genet* **31:** 224. doi:10.1016/J.TIG.2015.02.009

Stevens L, Félix MA, Beltran T, Braendle C, Caurcel C, Fausett S, Fitch D, Frézal L, Gosse C, Kaur T, et al. 2019. Comparative genomics of 10 new *Caenorhabditis* species. *Evol Lett* **3:** 217–236. doi:10.1002/EVL3.110

Sultanov D, Hochwagen A. 2022. Varying strength of selection contributes to the intragenomic diversity of rRNA genes. *Nat Commun* **13:** 7245. doi:10.1038/s41467-022-34989-w

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34:** W609–W612. doi:10.1093/NAR/GKL315

Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, et al. 2021. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* **49:** D605–D612. doi:10.1093/NAR/GKAA1074

Tanaka R, Okumura E, Kanzaki N, Yoshiga T. 2012. Low survivorship of dauer larva in the nematode *Caenorhabditis japonica*, a potential comparative system for a model organism, *C. elegans*. *Exp Gerontol* **47:** 388–393. doi:10.1016/J.EXGER.2012.03.001

Thomas JH. 2006. Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. *Genome Res* **16:** 1017–1030. doi:10.1101/GR.5089806

Thomas JH, Kelly JL, Robertson HM, Ly K, Swanson WJ. 2005. Adaptive evolution in the SRZ chemoreceptor families of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Proc Natl Acad Sci* **102:** 4476–4481. doi:10.1073/PNAS.0406469102

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14:** 178–192. doi:10.1093/BIB/BBS017

Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13:** 36–46. doi:10.1038/NRG3117

Vlaar LE, Bertran A, Rahimi M, Dong L, Kammenga JE, Helder J, Goverse A, Bouwmeester HJ. 2021. On the role of dauer in the adaptation of nematodes to a parasitic lifestyle. *Parasit Vectors* **14:** 554. doi:10.1186/S13071-021-04953-6

Volkening JD, Bailey DJ, Rose CM, Grimsrud PA, Howes-Podoll M, Venkateshwaran M, Westphall MS, Ané JM, Coon JJ, Sussman MR. 2012. A proteogenomic survey of the *Medicago truncatula* genome. *Mol Cell Proteomics* **11:** 933–944. doi:10.1074/MCP.M112.019471

Volkers RJM, Snoek LB, Hubar CJ, Coopman R, Chen W, Yang W, Sterken MG, Schulenburg H, Braeckman BP, Kammenga JE. 2013. Gene-environment and protein-degradation signatures characterize genomic and phenotypic diversity in wild *Caenorhabditis elegans* populations. *BMC Biol* **11:** 93. doi:10.1186/1741-7007-11-93

Wadsworth WG, Riddle DL. 1989. Developmental regulation of energy metabolism in *Caenorhabditis elegans*. *Dev Biol* **132:** 167–173. doi:10.1016/0012-1606(89)90214-5

Weber KP, De S, Kozarewa I, Turner DJ, Madan Babu M, de Bono M. 2010. Whole genome sequencing highlights genetic changes associated with laboratory domestication of *C. elegans*. *PLoS One* **5:** e13922. doi:10.1371/JOURNAL.PONE.0013922

Weinstein DJ, Allen SE, Lau MCY, Erasmus M, Asalone KC, Walters-Conte K, Deikus G, Sebra R, Borgonie G, van Heerden E, et al. 2019. The genome of a subterrestrial nematode reveals adaptations to heat. *Nat Commun* **10:** 5268. doi:10.1038/S41467-019-13245-8

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168:** 1041–1051. doi:10.1534/GENETICS.104.031153

Woodruff JB, Wueseke O, Hyman AA. 2014. Pericentriolar material structure and dynamics. *Philos Trans R Soc Lond B Biol Sci* **369:** 20130459. doi:10.1098/RSTB.2013.0459

Wu JY, Xiao JF, Wang LP, Zhong J, Yin HY, Wu SX, Zhang Z, Yu J. 2013. Systematic analysis of intron size and abundance parameters in diverse lineages. *Sci China Life Sci* **56:** 968–974. doi:10.1007/S11427-013-4540-Y

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586–1591. doi:10.1093/MOLBEV/MSM088

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22:** 1107–1118. doi:10.1093/MOLBEV/MSI097

Yednock BK, Neigel JE. 2014. Detecting selection in the blue crab, *Callinectes sapidus*, using DNA sequence data from multiple nuclear protein-coding genes. *PLoS One* **9:** e99081. doi:10.1371/JOURNAL.PONE.0099081

Zhang R, Roostalu J, Surrey T, Nogales E. 2017. Structural insight into TPX2-stimulated microtubule assembly. *eLife* **6:** e30959. doi:10.7554/eLife.30959

Received June 28, 2022; accepted in revised form November 18, 2022.