

**Proteome-wide positive selection analysis on
improved nematode gene annotation by
machine learning assisted
proteotranscriptomics**

Dissertation

Zur Erlangung des Grades
Doktor der Naturwissenschaften

Fachbereich Biologie
der Johannes Gutenberg-Universität Mainz

Camilo Alejandro Ceron Noriega

Mainz, 2023

Dekan:

1. Berichterstatter:
2. Berichterstatter:

Tag der mündlichen Prüfung: 20.03.2023

Contents

1	Summary	7
2	Zusammenfassung	9
3	List of publications	11
4	Aims of the Thesis	13
4.1	Nematode biology	14
4.2	Genome evolution and environment	16
4.3	The double-stranded DNA-binding proteins <i>tebp-1</i> and <i>tebp-2</i> form a telomeric complex with POT-1	17
4.4	Sources of genome assembly artifact: The beginning of error propagation	18
4.5	Challenges in genome annotation: Implications of high-throughput next-generation sequencing	20
4.6	Proteotranscriptomics	21
4.7	Machine-learning assisted Proteotranscriptomics data integration allows genome-wide annotation of protein-coding genes in 12 Rhabditea nematode species	22
4.8	Quantification of positive selection’s role as a driving evolutionary force	23
4.9	Different models	26
4.10	Article I: The double-stranded DNA-binding proteins TEBP-1 and TEBP-2 form a telomeric complex with POT-1	31
4.10.1	Summary	32
4.10.2	Zusammenfassung	33
4.10.3	Statement of Contribution	34
4.10.4	Article I: Main Text	36
4.10.5	Article I: Supplementary information	57
4.11	Article II: Nematode gene annotation by machine-learning-assisted proteotranscriptomics enables proteome-wide evolutionary analysis	75
4.11.1	Summary	76
4.11.2	Zusammenfassung	77
4.11.3	Statement of Contribution	78
4.11.4	Article II: Main Text	80
4.11.5	Article II: Supplementary information	99

4.12	Article III: AlexandrusPS - a user-friendly pipeline for genome-wide positive selection analysis	115
4.12.1	Summary	116
4.12.2	Zusammenfassung	117
4.12.3	Statement of Contribution	118
4.12.4	Article III: Main Text	120
4.12.5	Article III: Supplementary information	125
5	Discussion	141
5.1	Biases related to the reproductive mode in nematodes	141
5.2	Enrichments of networks related to cell division and spindle organization in the <i>Caenorhabditis</i> genus	142
5.3	<i>C. elegans</i> genes with positive selection evidence related to muscle function	143
5.4	Bias on positive selection detection with high divergent species	144
6	Conclusions	145
7	Acknowledgements	146
8	References	146
9	Curriculum Vitae	158

1 Summary

Numerous studies in the nematode and model species *Caenorhabditis elegans* have led to significant discoveries in the realms of biology and biomedicine. These are difficult to extrapolate in the context of evolution as the nematode phylum comprises considerably large phylogenetic diversity. Exemplifying this issue we tried to recapitulate the evolutionary history of the *tebp-1* and *tebp-2* genes which we showed to play significant roles in telomere biology in *C. elegans* (Article I). We could show that *Caenorhabditis briggsae* homologs of these proteins also bind telomeres. Expanding the evolutionary analysis by looking into phylogeny and synteny in eight additional *Caenorhabditis* nematodes we showed that these proteins may have a conserved role across the *Caenorhabditis* genus. Aiming for the detection of signs of positive selection we noticed lacking or inadequate gene annotations for many of the target nematodes. The accuracy of a positive selection analysis was also hampered by the quality of these gene annotations available in the **WormBase** nematode information resource which vastly depends on automated annotation workflows using available assembled genomes.

To close this gap, we employed a proteotranscriptomics technique along with machine learning-assisted quality control to improve gene annotations for 12 nematode species, enabling systems analysis and new insights into evolutionary processes (Article II). By comparing our method to the very elaborated *C. elegans* annotations, we demonstrated the system's power and identified 2 previously unidentified genes in this species (authorized by **WormBase** curators), which is remarkable after more than 20 years of diligent manual curation in this species. With our technique, we were able to produce high-quality annotations for 9 genome-sequenced species and provide new protein-coding gene annotations for 3 additional species without genome assemblies (*C. droshophilae*, *R. regina*, and *R. axei*) that are of the same quality as *C. elegans*.

To benchmark the annotations and facilitate evolutionary analysis we assembled a pipeline enabling orthology predictions and positive selection analysis. Implementing the pipeline enabled the detection of 23,090 orthologous groups across the proteotranscriptomics annotation of coding genes of the 12 nematode species. Using the pipeline for large-scale positive selection analysis we detected orthology groups under positive selection. Encouraged by these results we realized the benefit of the pipeline for the scientific community and will provide it publicly under the name

AlexandrusPS as a Docker image. AlexandrusPS allows users to run CodeML protocols on a desktop computer in an automated parallel manner, facilitating high-throughput analyses with no need for high-performance computer systems. The pipeline will be introduced to the community via an application note article in one of the bigger bioinformatics journals (Article III).

2 Zusammenfassung

Zahlreiche Studien an der Nematode und Modellspezies *Caenorhabditis elegans* haben zu bedeutenden Entdeckungen in den Bereichen Biologie und Biomedizin geführt. Diese sind im Kontext der Evolution schwer zu extrapolieren, da das Nematoden-Phylum eine beträchtlich große phylogenetische Vielfalt aufweist. Als Beispiel für dieses Problem haben wir versucht, die Evolutionsgeschichte der *tebp-1*- und *tebp-2* Gene zu rekapitulieren, von denen wir gezeigt haben, dass sie eine bedeutende Rolle in der Telomerbiologie in *C. elegans* spielen (Artikel I). Wir konnten zeigen, dass *Caenorhabditis briggsae* Homologe dieser Proteine ebenfalls Telomere binden. Indem wir die evolutionäre Analyse durch Untersuchung der Phylogenie und Syntenie auf acht weitere *Caenorhabditis*-Nematoden ausweiteten, zeigten wir, dass diese Proteine möglicherweise eine konservierte Rolle in der gesamten *Caenorhabditis*-Gattung spielen. Mit dem Ziel, Anzeichen einer positiven Selektion zu erkennen, stellten wir bei vielen der Zielnematoden fehlende oder unzureichende Genannotationen fest. Die Genauigkeit der positiven Selektionsanalyse wird durch die Qualität der in der WormBase Nematoden-Informationsressource verfügbaren Genannotation beeinträchtigt, die in hohem Maße von automatisierten Annotationsworkflows unter Verwendung verfügbarer sequenzierter Genome abhängt.

Zur Schließung dieser Lücke setzten wir eine Proteotranskriptomik-Technik zusammen mit einer durch maschinelles Lernen unterstützten Qualitätskontrolle ein, um die Genannotationen für 12 Nematodenarten zu verbessern, was eine Systemanalyse und neue Einblicke in evolutionäre Prozesse ermöglicht (Artikel II). Durch den Vergleich unserer Annotation mit der sehr guten Annotation von *C. elegans* demonstrierten wir die Leistungsfähigkeit unserer Methode und identifizierten 2 zuvor nicht identifizierte Gene in dieser Spezies (autorisiert von WormBase Kuratoren), was nach mehr als 20 Jahren sorgfältiger manueller Annotation bemerkenswert ist. Mit unserer Technik konnten wir qualitativ hochwertige Annotationen für 9 genomsequenzierte Arten erstellen und neue proteinkodierende Genannotationen für 3 weitere Arten ohne sequenzierte Genome (*C. droshophilae*, *R. regina* und *R. axei*) in der gleichen Qualität wie die von *C. elegans* bereitstellen.

Um die Annotationen zu benchmarken und die evolutionäre Analyse zu erleichtern, haben wir eine Pipeline erstellt, die Orthologievorhersagen und positive Selektionsanalysen ermöglicht. Die Implementierung der Pipeline ermöglichte die Bestimmung von 23.090 orthologen Gruppen, die die proteotranskriptomische Annotation der protein-kodierenden Gene der 12 Ne-

matodenarten umfassen. Unter Verwendung der Pipeline für umfassende positive Selektionsanalysen haben wir Orthologiegruppen unter positiver Selektion entdeckt. Ermutigt durch diese Ergebnisse haben wir den Nutzen der Pipeline für die wissenschaftliche Gemeinschaft erkannt und werden sie der Allgemeinheit unter dem Namen **AlexandrusPS** als Docker-Image zur Verfügung stellen. **AlexandrusPS** ermöglicht es Benutzern, **CodeML**-Protokolle automatisiert parallel auf einem Desktop-Computer auszuführen, was Analysen mit hohem Durchsatz ermöglicht, ohne dass Hochleistungs-Computersysteme erforderlich sind. Die Pipeline wird der Community über einen Application Note-Artikel in einer der größeren Bioinformatik-Fachzeitschriften vorgestellt werden (Artikel III).

3 List of publications

- Article I: The double-stranded DNA-binding proteins TEBP-1 and TEBP-2 form a telomeric complex with POT-1
- Article II: Nematode gene annotation by machine-learning-assisted proteotranscriptomics enables proteome-wide evolutionary analysis
- Article III: AlexandrusPS - a user-friendly pipeline for genome-wide positive selection analysis

Abbreviations

Nomenclature

ω dN/dS

AA Amino acids

BEB Bayes empirical Bayes

BM Branch model

BSM Branch-site model

dN Nonsynonymous substitutions per site

dS Synonymous substitutions per site

iHS Integrated Haplotype Score

INSDC International Nucleotide Sequence Database Collaboration

LRH Long-Range Haplotype

LRTs Likelihood ratio tests

ML Maximum likelihood

NGS Next Generation Sequencing

ORFs Open reading frames

PAML Phylogenetic analysis by Maximum Likelihood

PCM Pericentriolar material

PTA Proteo-Transcriptomics Assembly

RNA-seq RNA sequencing

SSM Site-specific models

4 Aims of the Thesis

In an omics-level effort, I demonstrated the efficacy of a proteotranscriptomics approach combining high-throughput experimental data (RNA-seq and peptide evidence) for accurate protein-coding gene annotation in 12 nematode species including animals without a reference genome. This led to highly reliable and experimentally validated annotations which improved currently available annotations but also enabled the annotation of species without genome assemblies. In my article [1], I exemplify the potential of the resulting dataset to significantly advance evolutionary studies. These included genome-wide positive selection analysis, which led to the development of the pipeline **AlexandrusPS** (Article III) enabling comprehensive evolutionary analysis of any species to understand their genomic adaptations to their environments.

4.1 Nematode biology

Caenorhabditis elegans is a nematode species that has become a highly valuable model organism in scientific research due to its numerous advantages, such as its simple anatomy, fast reproduction, ease of cultivation, transparency, and complex cellular physiology [2]. Being the first metazoan to be fully sequenced [3], *C. elegans* has a well-annotated genome with 20,140 protein-coding genes, making it useful for studying animal genomes and development. Its simple anatomy and fast reproduction have made *C. elegans* particularly valuable for investigating various biological processes, and it has proven especially useful for studying molecular and cellular pathways shared with mammals, including those associated with human diseases [2]. Interesting evolutionary patterns and 60-80% orthology conservation with humans make it useful for studying the effects of aging and diseases caused by gene mutations. Characterized by simple general anatomy and basic nervous system, it is a valuable tool for investigating complex biological processes such as neurodegenerative diseases and functional synapses[4]. These advantages have made *C. elegans* a valuable resource for comparative biology and genetic research, leading to its widespread use in the field of medicine as a model [2] [5] [6].

C. elegans has also gained recognition for its simple reproductive system and its capability to yield rapid results and thus is a valuable model organism in the pharmaceutical industry. Its versatility in drug development includes target validation, lead optimization, and toxicity assessment. Despite its limitations in directly predicting drug action in humans, *C. elegans* provides important insights into human diseases through studies of specific molecular pathways and pharmacological approaches [2].

In the field of fertility and reproductive health, *C. elegans* has garnered recognition primarily due to the advantages presented by its straightforward reproductive system. The unique reproductive biology of *C. elegans*, characterized by self-fertilizing hermaphrodites and facultative males, results in a reduced genetic diversity and a smaller effective population size. Furthermore, the *C. elegans* hermaphrodite gonad is widely used in developmental and cell biology studies, making it a valuable research tool. The study of vulva development serves as a model system for investigating signaling processes in animal development and provides insights into evolutionary developmental biology. Hence, over the past four decades, research in *C. elegans* has played a critical role in advancing our understanding of development, neurobiology, biomedical research, and evolutionary biology

[2] [5] [6].

Research in related species play an important role in extrapolating information from *C. elegans* in an evolutionary context. This is because comparative analysis of closely related species can enhance our understanding of life history, genomic, and morphological evolution. Phylogenetic analysis is critical for accurate extrapolation and gene identification [7]. This information can then be used to determine ancestral and derived states of characters, which is crucial for comprehending the evolution of the lineage. The genus *Caenorhabditis*, of which *C. elegans* is a member, has undergone significant speciation and diversification over at least 30 million years, presenting a challenge in determining homology between species. Nevertheless, evolutionary studies in *Caenorhabditis* can offer deeper insights into the evolution of particular lineages compared to more distantly related organisms [4] [5]. Despite the genus' limited morphological diversity compared to that found among vertebrates, it is characterized by a high degree of genetic variation. The genus includes species that exhibit two distinct modes of reproduction: self-fertile hermaphrodites and outcrossing species. Self-fertile hermaphrodites are widespread in the genus and evolved independently multiple times across the nematode phylogeny with varying mechanisms[5]. Genomic comparisons between these two modes show that outcrossing species have larger genomes and high polymorphism than self-fertile hermaphrodite species, indicating that the transition from outcrossing to self-fertilization resulted in a decrease in genome size and altered molecular evolution patterns affecting both coding and non-coding regions [4] [6].

At the genome level *Caenorhabditis* species display high degrees of conservation in protein-coding exons, with highly conserved sequence elements found within coding exons in *C. elegans*, *C. inopinata*, and *C. briggsae* [6][5]. However, *Caenorhabditis* genomes also exhibit a high rate of intron turnover, characterized by frequent intron loss and rare intron gain[4]. An evolutionary explanation for this pattern is high genetic drift in small populations. Differences in genome size among *Caenorhabditis* species may be due to ecological changes associated with variations in life history, however, this hypothesis can only be tested through comparisons with more evolutionarily divergent species such as members of the Rhabditida family [6] [7].

Rhabditidae are a prominent taxonomic unit in the field of evolutionary biology, encompassing over 200 described species, including the *Caenorhabditis* genus. Its origin dates back to approximately 750 - 650 million years

ago [8]. Members of this taxonomic group are crucial for genetic and molecular analysis, with *Oscheius tipulae* and *Pristionchus pacificus* playing a central role in this respect. Specifically, *P. pacificus* has made invaluable contributions to our understanding of vulva development in nematodes, advancing our knowledge of the molecular mechanisms involved beyond the previously well-studied species *C. elegans* [9]. Another notable species in this family is *Panagrellus redivivus*, which was the first to be compared with *C. elegans* at the cell lineage level [10]. Comparative analysis of the diverse species within the Rhabditida family researchers promises the discovery of new genetic and regulatory features, furthering the understanding of evolutionary processes. To enable such comparative studies **WormBase: Nematode Information Resource** [11] was established. **WormBase** is a comprehensive resource for nematode information, providing extensive data on genomes, gene models, genetics, mutant and RNAi phenotypes, expression, interactions, literature, and data-mining tools. As a vital toolbox for the study of biological phenomena in *C. elegans* and other nematodes, **WormBase** plays a crucial role in advancing our understanding of this model organism within an evolutionary context [2].

4.2 Genome evolution and environment

Protein-coding CDS are the DNA sequences required for the synthesis of functional proteins. CDS comprise nucleotide triplets known as codons, which are translated into amino acids (AA) during protein production according to the genetic code of each organism [12]. The genetic code is known as a partially redundant system because several codons encode the same amino acid. This redundancy is manifested by synonymous sites, for which specific changes in the coding DNA sequence do not change the amino acid sequence and thus structure or function of the protein remain unchanged. In the absence of selective forces, beneficial mutations may be selected and fixed via drift effects [13]. Conversely, mutations that encode distinct amino acids (known as non-synonymous sites) might be selected against and disappear from the genome since they are not favorable. The selective forces that act on orthologous proteins (proteins from distinct species that descended from a common ancestor) are important drivers of molecular evolution [14].

For decades, the rates and patterns of molecular sequence evolution have been estimated using comparative studies of orthologous genes [15]. These studies have shown that in the protein-coding regions of these genes two types of sites have evolved differently [8]. The observation of conserved

amino acid sequences with synonymous changes in the genomic code suggests genetic drift, which refers to the alteration in the frequency of an existing gene variant in a population as a result of random processes . On the other hand, highly variable sites signify genomic locations that were shaped by natural selection, such as purifying or positive Darwinian selection [16] [17]. Studies of such sites can reveal how genomes evolve, providing researchers with insights into the biological importance of genes of interest and species differences [18].

Protein-coding gene evolution is driven not just by mutation, but also by other mechanisms such as alternative splicing of exons or introns, which increases diversity in these codons [19]. Such modifications may have an evolutionary impact on mutation rates and amino acid exchange, which has long been a driving force in protein evolution [20]. Thus, coding-sequence evolution is the outcome of mutational processes interacting with molecular selection forces which might confer a fitness benefit and result in adaptive evolutionary diversification of proteins [13] [21].

Studies of the adaptive diversification of proteins have shown that the environment has a great impact on the evolution of organisms [22] [23] [24]. In these studies, researchers used a variety of methods, including phylogenetic analysis and model-based methods, to identify selective pressures on coding regions. Quantifying the mode and degree of selective pressure on coding regions can provide insight into how orthologous genes evolved differently in different species in response to different environments [25].

4.3 The double-stranded DNA-binding proteins *tebp-1* and *tebp-2* form a telomeric complex with POT-1

In this study, our aim was to discover and describe previously unknown telomere-associated proteins in *Caenorhabditis elegans*. To achieve this, we combined quantitative proteomics with a DNA pulldown experiment performed with nuclear extract from *C. elegans*. This allowed for the identification of proteins that associate with telomeres. Our findings suggest the existence of a previously undescribed complex of telomere-binding proteins in *C. elegans*. From this group, we focused on the *teb*paralogs R06A4.2 and T12E12.3 and conducted further experiments to determine their roles at *C. elegans* telomeres.

An evolutionary analysis of the *tebp* family verified that these genes are present only in the *Caenorhabditis* genus, mostly in the *elegans* super-

group, and that the number of protein-coding *tebp* genes varies per species. In our study, we observed a high degree of regional synteny conservation between the *tebp-1* gene of *C. elegans* and several other species within the *Caenorhabditis* group. However, *tebp-2* did not show any signs of regional synteny across *Caenorhabditis* species, suggesting that the gene duplication event creating *tebp-2* occurred after the divergence from the *C. inopinata* lineage which is a close sister species of *C. elegans*. Finally, the study also found that CBG11106, the single homolog of *tebp-1* and *tebp-2* in *C. briggsae*, can also bind to telomeric DNA, suggesting that *tebp* proteins are general telomere-binders in the *elegans* supergroup. In our study, we observed that the N-terminal region of *tebp* genes displays greater similarity between orthologs in comparison to the C-terminal region, which presents challenges for correct alignment. The utilization of an online tool like Datamonkey [26] to identify signals of positive selection was rather unsatisfactory. This was not only attributed to the limited functionality of the program but also to problems with available annotations. We observed that the quality of genome assemblies and annotations of species outside of the *Caenorhabditis* group of nematodes posed significant challenges for our evolutionary analyses. This resulted in substantial difficulties in accurately analyzing evolutionary patterns and trends in these species.

4.4 Sources of genome assembly artifact: The beginning of error propagation

A large number of new genome sequence data has become available due to the rise of the genomic era, advances in Next Generation Sequencing (NGS) technologies, and ongoing genome sequencing projects such as the Genome 10K project [27] or the high-resolution map of human evolutionary constraint using 29 mammals [28]. This availability enables researchers to conduct comparative analyses across many species [29] [30].

Using such comparative analyses researchers identify and characterize genetic changes underlying phenotype differences between species [31] [32]. Genomes are highly variable at the nucleotide level, and this variation has been used to study both within-species diversity and to identify patterns of evolution [33].

One approach to identifying patterns of evolution in comparative analysis is to test evolutionary hypotheses to measure selective pressures. Such analyses include investigating positive selection to explain why some genes are more common in a population than in others. It also offers an explanation for why some species have evolved new traits after their environment

has changed a process also called trait decay. ([34] ; [35], [36],-). In addition, evolutionary inferences can be used to guide the planning of functional and validation studies related to specific traits, behaviors or ecotypes (e.g., phenotypic plasticity) [37]. Such analyses may also be used to study the genomic basis of adaptive evolution as well as assess biological significance [38].

Genome sequences evolve in response to a wide range of interacting evolutionary forces, rather than according to a phylogenetic model of sequence evolution [39]. As a result, any technical error introduced during genome assembly or annotation will mask the effect of all these evolutionary processes, rendering any assumptions made by the inference model meaningless.

The rapid increase in genomic sequencing due to decreasing costs has made genomic data a vital resource for biotechnology and medical research. The quality of genome assembly is a critical factor in understanding the genetic makeup of species and the accuracy of gene function annotations [40]. However, repeat structures and heterozygosity in genomes can pose challenges in genome assembly, leading to decreased completeness, missing genes by gaps and inaccuracies in genome size estimation [41]. False duplications of genomic information are a common issue in genome assembly, resulting from sequencing errors and higher heterozygosity. These duplications come in two forms: heterotype and homotype, with heterotype duplications occurring in regions with greater sequence divergence between paternal and maternal haplotypes and homotype duplications stemming from sequencing errors. Using long-read data can reduce the frequency of false duplications, providing increased accuracy and comprehensiveness and improving assembly quality [41]. Additionally, genomic repeat content still affects genome assemblies, leading to an increase in gaps and collapsed regions, unresolved segmental duplications, and a smaller proportion of high-copy repeats. [42]. Another challenge in genome sequencing and assembly is the contamination from sources such as bacteria, sequencing vectors, or human DNA [43] [42] [41] [44], exacerbating the situation.

The International Nucleotide Sequence Database Collaboration (INSDC) is the largest repository of whole-genome DNA sequence information for eukaryote species. As of March 2021, it contains data for 6,480 unique species, but only 583 of these (9%) have reference-quality, chromosome-scale assemblies. The rest are draft assemblies that do not meet current quality requirements. Although there are plans to significantly increase

the number of reference-quality genomes, this will require the efforts of the community or large consortia, and may still take some time.

Genome annotation is the process of identifying functional elements in genomic sequences, and it is based on accurate and contiguous genomic sequences. The first sequenced organisms underwent a highly curated annotation process, but with the large number of sequenced genomes today, automated processes are now needed for efficiency. These processes screen the genome sequence for open reading frames that potentially code for proteins. While this enables efficient prediction of possible open reading frames, the accuracy has been reported to suffer in many cases.

4.5 Challenges in genome annotation: Implications of high-throughput next-generation sequencing

The exponential growth in genomics, driven by the decline in sequencing costs and the accumulation of genomic data, has resulted in a substantial increase in genome information. However, integrating functional and structural information into a genome sequence, known as genome annotation, remains challenging because of the potential for limited accuracy. To address this challenge, automated genome annotation uses computational methods to identify features such as open reading frames within the genome sequence [43]. The process of genome annotation, which can identify genes, their functions, non-coding RNAs, enhancer sequences, and methylation sites, is crucial for downstream applications and understanding the functional and evolutionary aspects of genomes [44]. Automated genome annotation pipelines use the genome sequence to predict open reading frames (ORFs) using *ab initio* gene-finding methods and the alignment of evidence at the level of homology, EST and functional domains for validation.

The annotation of large and fragmented genomes presents a significant challenge, as it is prone to errors and contamination in draft assemblies [45]. Hence, even a very well functioning annotation pipeline will not be able to annotate highly fragmented genomes in a satisfactory manner. Limitations in annotation pipelines, such as differences in quality and operator expertise, can cause the propagation of errors across species and negatively affect the precision of gene annotation[46]. These limitations can cause a range of errors mainly due to incorrect parameters in pipelines, and lead to missed genes, mispredictions, inaccurate gene naming, distinct names for the same gene, and mistakes in genomic coordinates and assigned func-

tions [47]. The propagation of errors in existing gene annotations can have a cascading impact on related species. Correction of a single annotation error requires adjustments to all dependent annotations [44]. This is even worse in light of some automated gene annotation pipelines that primarily use sequence information as a basis for prediction, without additional evidence to support their accuracy [43].

Accurate annotation of a newly sequenced organism's genome is essential for obtaining a comprehensive understanding of its genome and its unique traits. The significance of precise gene annotation cannot be overstated, as inaccurate annotations can negatively affect evolutionary studies and biological comprehension [47]. The accuracy of gene annotation is a critical factor in genomics, and improving the precision of annotation pipelines and annotations in relevant databases is crucial for advancing our understanding of genomes [45].

4.6 Proteotranscriptomics

The utilization of high-throughput experimental data, such as RNA sequencing (RNA-seq) and proteomics, provides a reliable method for gene annotation improvement, leading to greater accuracy in results. The rapid progress in sequencing technology has made transcriptome assembly a popular and cost-effective alternative approach for gene prediction. In this process similar to genome assembly the cDNA sequence reads can be assembled into transcripts. The approach was strongly enforced by the development of dedicated transcriptome assembly programs. Combined with the prediction of potential open reading frames it has developed into a valuable tool for gene discovery across a variety of organisms. The increased prevalence of transcriptome studies in recent years reflects this trend [48] [49] [50] [51] [52] [53] [54] [55] [56] [57].

Transcriptome assembly is a crucial step in the analysis of gene expression and can be performed using either a genome-free or genome-guided approach. The genome-guided assembly aligns the reads to a reference genome, partitions the reads based on the locus, and then performs de novo assembly at each locus to reconstruct transcripts. The genome-free assembly does not use a reference genome and instead reconstructs transcripts solely from the actual read sequences. The choice between the two approaches depends on the research question, with genome-guided assembly being useful for capturing sequence variations when the sample's genome differs from the reference genome. In contrast, genome-free assem-

bly is preferred when the genome assembly is missing or of poor quality and employs either overlap graph or de Bruijn graph algorithms, with the latter being widely used in RNA-Seq assemblers due to its lower computational complexity.

Despite its advantages, genome-free transcriptome assembly has limitations, such as assembly errors, chimeras, and misestimation of allelic diversity. Long-read sequences can enhance accuracy, however, they also introduce new errors. Hybrid approaches combining short and long-read sequences, along with quality control measures and benchmarking programs, have been proposed to mitigate these limitations [43]. ORF prediction from transcriptome assemblies can be improved by incorporating reference genome annotations available in current databases. These transcriptome assemblies provide valuable insights into novel genes, alternative splicing, and inform functional genomics, comparative genomics, and evolutionary patterns in non-model organisms. They serve as an important resource for identifying gene functions and evolutionary patterns.

In order to validate and increase confidence in transcriptome predictions, peptides can be employed as evidence. The utilization of mass spectrometry in cross-validating predicted ORFs with peptide identifications eliminates misassembled transcripts and enhances confidence in protein predictions. This process not only enables expression estimation and sequence variant identification but also facilitates the annotation of high-confidence open reading frames. The Proteo-Transcriptomics Assembly (PTA) approach integrates transcriptome assembly and peptide evidence to enhance gene discovery, comparative analysis, and ORF annotation. De novo transcriptome assembly forms the basis for the unbiased PTA approach, making it applicable to species without prior genome information. The addition of peptide evidence to transcriptome assembly predictions results in an increased proportion of complete transcripts and improved accuracy in gene annotations.

4.7 Machine-learning assisted Proteotranscriptomics data integration allows genome-wide annotation of protein-coding genes in 12 Rhabditea nematode species

The challenge of obtaining high-quality gene annotations in nematodes has persisted, however, we implemented the integration of a cutting-edge proteotranscriptomics approach with machine learning quality control as a highly effective solution to bridge this gap. High-quality protein-coding gene annotations were generated for 12 nematode species, including species

outside of the *Caenorhabditis* lineage. The effectiveness of this approach was demonstrated by the discovery of 2 previously unknown genes in the well-annotated *C. elegans* and a prediction rate of over 90% of the 20,127 *C. elegans* WormBase gene models. Additionally, the study emphasized the marginal accuracy of some of the previously established annotations, e.g. it identified hundreds of falsely merged genes in the widely used nematode *P. pacificus*.

In addition, we used our versatile pipeline to provide annotations for three species whose genomes have not been previously sequenced or assembled - *C. drosophilae*, *R. regina*, and *R. axei* -thereby rendering it a useful resource for evolutionary studies. Orthology analysis of 23,090 groups among 12 nematode species was conducted, followed by a rigorous positive selection analysis of up to 5,400 orthology groups, complemented by an enrichment analysis to unravel adaptive mechanisms in specific gene families and pathways. The results imply that nematode species have undergone evolution to enhance their adaptation to their surroundings through changes in genes involved in stress response, detoxification, metabolism, reproduction, and development. This study shows the power of the technique and the resulting dataset for evolutionary analyses across a broad phylogeny and offers a valuable foundation for future evolutionary proteomic investigations.

4.8 Quantification of positive selection's role as a driving evolutionary force

Orthologous sequences have been modified by an extensive evolutionary process with mutation rates that vary by orders of magnitude [58]. Accounting for rate variation under different levels of selective pressure would thus provide insight into the functional restrictions on proteins. Proteins with strict functional or structural requirements face significant purifying (negative) selective pressure, resulting in fewer amino acid modifications. Consequently, genes with limited rate of evolution are prone to perform critical functions optimally. Unless their interaction networks are altered, the probability of improved performance is relatively low. Genes having redundant and non-central functions, as well as weaker constraints, evolve at a faster rate. A certain fraction of these genes may have been subjected to recent positive selection. This could be because the evolutionary rate of protein-coding genes is affected by their dispensability, which affects the rate of evolution. [21] [59].

Positive selection-detecting statistical models of molecular evolution are

viable tools to investigate such processes. These models can be classified into two broad types, each of which is better suited to investigating processes at different time scales. [60]. Long-Range Haplotype (LRH) and the integrated Haplotype Score (iHS) [61], [62] examines positive selection that has recently occurred among populations [63] [61]. The other type, known as codon models, detects positive selection between species; it is more suited to inferring earlier events, such as species divergence. [64] [65]. Polymorphism and divergence data can be used by methods from any class [66]. Positive selection signals may be present across species but not in populations, or vice versa.

The relative contributions of positive selection and neutral drift to the evolution of a set of orthologous genes can be determined by measuring their respective evolutionary rates. This measuring is computed by comparing the estimation of substitution rates and selective constraints in coding regions among a group of orthologous genes' multiple sequence alignments. This calculation can be deduced by estimating the excess of nonsynonymous substitutions per site (dN), as compared to synonymous substitutions per site (dS), known as the dN/dS ratio or ω [19] [67] [68].

The dN/dS ratio (ω) is thus the proportion of amino acid-altering, nonsense or missense mutations (non-synonymous) out of silent mutations (synonymous) [69]. ω determines whether certain sites in the genome have been subjected to positive or purifying selection. For example, if a mutation in a gene conferring an advantage occurred and became more common in subsequent generations, ω can infer which selective pressures produced this shift [70].

The calculation of the selection pressure on the coding sequence in a neutral evolutionary scenario can be done on the premise that synonymous mutations accumulate neutrally. A site with ω less than 1 suggests purifying selection [71], while sites with ω equal to 1 suggest neutral evolution. Positive Darwinian selection occurs at the protein level and is shown by sites with a ratio value greater than 1; all such positively selected sites may be interpreted as occurred through molecular adaptation, conferring an evolutionary advantage to the organism [16].

Although ω provides a simple way of quantifying the number of modifications per site and thus revealing the selection pressure acting on several protein-coding regions, it is challenging to determine which genes or sites have been affected by adaptive evolution because only a few species or

sites are typically affected. The more likely interpretation is that positive selection has occurred at specific evolutionary branches and gene-specific sites, requiring testing of each possibility as a model to examine evolutionary processes in depth. Several computational methods for calculating selection pressure from orthologous proteins in a phylogenetic context have been developed. Over the last 30 years the most commonly used method for assessing selection pressure from protein-coding sequences has been a pair-wise comparison method of maximum likelihood (ML) codon-based models. This allows researchers to evaluate which pair-wise model comparisons best reflect orthologous gene molecular evolution [72]. This method is also implemented in Phylogenetic analysis by Maximum Likelihood (PAML) [73], resulting in a well-established workflow with a long history of successful usage. It is highly accurate and statistically robust for analyzing genome-wide data, documenting selection pressure on codon use in protein-coding genes, and evaluating evolutionary selection hypotheses [30]. CodeML, the main software in PAML, uses an empirical Bayes technique to identify codons undergoing adaptive evolution [13]. It detects positive selection by comparing pairs of nested statistical models via likelihood ratio tests (LRTs) to assess whether positive selection might have occurred. Various hypotheses can be tested by comparing these statistical models pair-wise and estimating adaptive selection along a coding gene's phylogeny using likelihood-based assessments [74]. CodeML also estimates parameters in models with changing rates among sites and several genes [75].

CodeML can evaluate selection signatures in two stages. The first stage comprises running different models with different assumptions regarding how ω varies across a multiple sequence alignment (MSA) and/or phylogeny. Three models in CodeML stand out from the rest to evaluate ω at various levels. The first focuses on changes of ω at various gene sites which remain constant across branches (site-specific models: SM) [23]. The second centers around the assumption that ω can change in various phylogenetic branches but remains constant across the coding sequence (branch model: BM) [76], and the third assumes ω can change at specific sites and in specific branches (branch-site model: BSM) [77] [78]. To test each of these models, the model, its parameters, and an orthologous gene coding-sequence alignment has to be provided in the control file along with a species tree [30] [79]. In the second stage, for all models, a Likelihood Ratio Test (LRT) is used to examine the goodness-of-fit between two nested models and determine which of these fits the dataset better.

4.9 Different models

- **Site-specific models (SM):** These allow ω to vary across MSA sites. The models include individual comparisons (M0 vs. M3, M1a vs. M2a, M7 vs. M8, M8a vs. M8 [80]) for using Darwinian selection's successive LRT. It has two types of site class-specific models: the first comprises alternative classes (models 3 (M3), 2 (M2a), and 8 (M8), the second includes null classes (models 0 (M0), 1 (M1), 7 (M7), and 8a (M8a)). The M1a and M2a models belong to the same model pair. M1a assumes genetic drift and fixes ω values to 1, whereas M2a allows for adaptive selection [81]. The second model pair is composed of the M7 and M8. M7 assumes that ω is beta-distributed among sites (interval 0,1) and thus excludes positively selected sites, whereas M8 introduces an additional class of sites that enables positive selection. [82]. The M1a vs. M2a comparison is regarded as less strong than the M7 vs. M8 comparison. The location is assumed to be positively selected if M2a or M8 are much more likely than M1a or M7. Furthermore, when LRTs are significant, CodeML computes the posterior probability of sites under selection using the Bayes empirical Bayes (BEB) test [82], which may identify specific sites under positive selection if the LRT of the M1a vs. M2a or M7 vs. M8 comparison is significant [13] [24].
- **Branch models (BM):** Due to the heterogeneity in evolutionary mechanisms between species and along sequences, estimating ω is difficult. To tackle this challenge, statistical phylogenetics with branch models is applied [83]. These models allow the ω of various phylogenetic branches to vary, enabling positive selection on specific lineages to be inferred by comparing their ω values (M2). The M2 LRT compares a null model that assumes that ω equals 1, implying neutral selection across the phylogeny, to a model that allows ω to fluctuate above 1. If the null model is less likely than its counterpart, positive selection pressure is inferred for that branch or node [13].
- **Branch-site models (BMS):** It allows ω to vary among sites along selected branches of a phylogeny, so that foreground lineages can be fitted with models with and without positive selection. Like branch

models, the null hypothesis of no positive selection can be tested with LRT to compare the likelihood of different models fitting the data better than one without it. The branch-site model compares the branch site with a null model under a fixed- ω assumption ($\omega = 1$). A BEB test calculates posterior probabilities of specific sites under positive selection. The branch-site models, like the branch models, rely on a set of identical trees with all foreground branches labeled. Because classifying each branch as foreground or background does not always rely on strong a priori information, it makes sense to test many or all branches in the tree, with each branch considered as the foreground branch in turn.

Despite CodeML enabling the investigation of all the above-mentioned models, the program executes on a single CPU and thus is not suited for large numbers of species or longer sequences. Additionally, the program also does not calculate P -values automatically for model likelihood comparisons. Its output might be confusing to inexperienced users because it does not emphasize or show the most important results in a straight-forward way. Furthermore, it requires the manual generation of unique configuration files for each alignment and test, complicating the high-throughput analysis of many sequences [13] [84] [85]. Another significant obstacle to be overcome is the amount of resulting output data. We grouped the user-required tasks into three categories of problems that CodeML users may experience: 1) collecting and organizing the input data, and modifying CodeML configuration files; 2) performing and compiling all required estimations; 3) results analysis, comparisons, and calculations. A thorough explanation of these categories can be found below.

- Collecting and organizing the input data, and modifying CodeML configuration files: This category refers to the tasks that users must perform before running CodeML. They include choosing orthologous groups for the positive selection analysis used for phylogenetic tree construction and for Multiple Sequence Alignment (MSA) of the relevant orthologs. The relevant files must be structured into directories that are accessible to CodeML, including user-specific configuration files that have been edited and are unique to each individual model that will be utilized.
- Performing and compiling all required estimations: This category describes the tasks that need to be accomplished during and after running CodeML. They include the collection of all ML parameter estimations from the output, which is not very straight-forward because it does not

provide a comprehensible display of key results.

- Results analysis, comparisons, and calculations: Finally, the user needs to estimate all LRT comparisons and compute P -values to determine positive selection after extracting all required ML parameters [24] [13].

There are many bioinformatics resources available to study evolutionary forces acting on a gene and resolve the discussed problems of CodeML. On one side, web servers have been created that implement different methods of hypothesis testing and ancestral sequence reconstruction using codon data. These include PSP [86], PhyleasProg [87], the SNAP server [88], HyPhy environment [89] and Selecton version 2.2 [16]. All involve SM, but PSP and PhyleasProg also allow BSM analyses. Besides Web-server implementations, a variety of desktop software packages are available for protein structure modeling. These tools can be divided into two categories: single-task (JCoDA [85], Armadillo [90], PAMLX [84], IMPACT_S [91]) and multi-task (IDEA [22], gCodeML [92], POTION [93], VESPA [94]). Single-task software allows users to perform SMs in all the software packages, while BM analyses are only available in IDEA, Armadillo and PAMLX and BSM are possible in gCodeML, VESPA, Armadillo, and PAMLX. These tools provide a major advance in data-intensive research. However, there are significant limitations: they are too complex to set up and configure [85], and usually require infrastructures with limited availability or enhanced informatics skills.

Despite this large number of tools for positive selection analysis, there still is a need for software that greatly reduces the manual input required for such analyses. Tackling this issue we have developed a computational pipeline called **AlexandrusPS**, which facilitates researchers in performing correct and efficient large-scale evolutionary analyses with any of the described codon substitution models (SM, BM, and BSM).

AlexandrusPS is a pipeline implemented as a combination of scripts written in different programming languages, including Perl [95], R [96], and shell [97]. It runs in a Linux/UNIX environment and consists of 19 Perl scripts and 3 R scripts called by the main shell script ‘**AlexandrusPS.sh**’. **AlexandrusPS** is designed to handle each step of the CodeML workflow, thus minimizing user intervention. The main functionalities of the pipeline are described in the attached manuscript draft named ‘**AlexandrusPS: a user-friendly pipeline for genome-wide positive selection analysis**’ which we in-

tend to submit to a peer-reviewed scientific journal. A detailed manual can be found on the **AlexandrusPS** GitHub page (<https://github.com/alejocn5/AlexandrusPS>).

AlexandrusPS is a PC-based pipeline that is packed in a Docker image to avoid the need for local installation of any modules or programs. The pipeline is provided as an open-source solution, allowing users to run various **CodeML** models for molecular adaptive evolution (SSM, BM, and BSM) in parallel. Based on standard protocols, **AlexandrusPS** can analyze large-scale, genome-wide datasets with default parameters and requires only the CDS and peptide FASTA files of the proteins of interest for input. With its user-friendly interface, **AlexandrusPS** offers significant advantages over other programs.

AlexandrusPS also solves two additional challenges in the steps needed before performing the actual positive selection analysis. These include accurate orthology predictions and sequence alignment. This is important as including ancient paralogs, i.e. paralogs that have diverged during long timescales has been shown to bias positive selection analysis [98]. The increased rate of nonsynonymous substitution caused by decreased purifying selective pressure can result in two alternative fates of the gene copies. Either one of the paralogs becomes non-functional due to the lack of selective pressure and accumulation of mutations. In some cases, the functions and expression patterns of the gene pair may diverge substantially and give rise to novel functions or specializations in the organism also called neofunctionalization [99]. In this case one copy may be under positive selection pressure, while the other copy may be under purifying selection. Including both copies in the same alignment and positive selection analysis can result in indecisive signals [100] [101]. Furthermore, the presence of sequence and alignment errors can hinder the accuracy of the positive selection detection process [102]. The robustness of a test for positive selection to these types of errors is challenging to develop [103]. Factors that limit alignment accuracy are high sequence divergence and differential evolutionary constraints across different structural components of a protein. In addition to these difficulties in aligning sequences, differences in codon usage patterns between species can increase false positives and render accurate positive selection detection challenging [102] [103]. The power of a test for positive selection is limited at both extremes of sequence divergence, with little inference power at low divergence and an overwhelming amount of synonymous substitutions resulting in alignment errors at high divergence [102]. Alignment errors are prevalent in many commonly used alignment programs and can result in false positive results in the detection of posi-

tive selection. These errors occur when non-homologous codons or amino acids are placed in the same alignment position. Such misalignments are mainly caused by ortholog misassociation and indels [104]. Among a range of programs, PRANK [105] has a significant advantage over other commonly used alignment programs, such as MUSCLE [106], MAFFT [107], and ClustalW [108], in accurately aligning sequences. This is because PRANK [105] takes evolutionary information into consideration during both codon alignment and gap placement. The importance and benefit of incorporating evolutionary information into sequence alignment is further highlighted by the consistently better performance of PRANK [105] compared to other programs [109].

AlexandrusPS automatically generates orthology relationships and identifies optimal orthology groups for positive selection analysis to avoid issues such as paralog introduction. It uses PRANK [105] to align relevant ortholog group sequences and also creates a gene tree of each OGC. AlexandrusPS then organizes, executes, and extracts all necessary information from CodeML outputs, fully automating the analysis process without any need for user intervention. The pipeline generates four main outputs: Orthology relationships, site-specific positive selection results, branch and branch-site positive selection results, along with all intermediate files for each OGC. These intermediate files enable manual repetition of specific analyses for individual OGCs without having to repeat the entire process. AlexandrusPS allows users to run CodeML protocols on desktop computers in an automated parallel manner, facilitating high-throughput analysis without the need for high-performance computing systems.

4.10 Article I: The double-stranded DNA-binding proteins TEBP-1 and TEBP-2 form a telomeric complex with POT-1

4.10.1 Summary

In this study, we aimed to identify and characterize telomeric factors in the model organism *Caenorhabditis elegans*. We utilized a quantitative proteomics approach in conjunction with a DNA pulldown experiment using *C. elegans* nuclear extract, which led to the identification of a set of proteins that associate with telomeres. Our focus was on the paralog proteins R06A4.2 and T12E12.3, and we further characterized their function at *C. elegans* telomeres.

Our results showed that *tebp* genes are present only in the *Caenorhabditis* genus, mainly in the *elegans* supergroup, with a variable number of protein-coding genes per species. Additionally, a single homolog of *tebp-1* and *tebp-2*, CBG11106, in *C. briggsae* was also found to bind telomeric DNA, suggesting a potential conserved telomere-binding function throughout the *Caenorhabditis* genus. Our data revealed a high degree of regional synteny conservation between the *tebp-1* gene of *C. elegans* and other species within the *Caenorhabditis* genus, while *tebp-2* showed no signs of regional synteny. Our findings provide evidence for the first described complex of telomere-binding proteins in *C. elegans*, including the first reliably described *C. elegans* telomere double-strand binders R06A4.2 and T12E12.3.

4.10.2 Zusammenfassung

Das Ziel unserer Studie war die Identifizierung und Charakterisierung von telomer-bindenden Faktoren im Modellorganismus *Caenorhabditis elegans*. Wir verwendeten einen quantitativen Proteomik-Ansatz in Verbindung mit einem DNA-Pulldown-Experiment unter Verwendung von *C. elegans* Kernextrakten, was zur Identifizierung einer Reihe von Proteinen führte, die mit Telomeren assoziieren. Unser Fokus lag auf den Paralog-Proteinen R06A4.2 und T12E12.3, deren Funktion an *C. elegans* Telomeren wir weiter charakterisierten.

Unsere Ergebnisse zeigten, dass *tebp*-Gene nur in der Gattung *Caenorhabditis* und hauptsächlich in der Elegans Supergruppe vorhanden und mit einer variablen Anzahl von proteinkodierenden Genen pro Art repräsentiert sind. Darüber hinaus wurde festgestellt, dass ein einziges Homolog von *tebp-1* und *tebp-2*, CBG11106, in *C. briggsae* auch telomerische DNA bindet, was auf eine potenziell konservierte Telomer-Bindungsfunktion in der gesamten Gattung *Caenorhabditis* hindeutet. Unsere Daten zeigten ein hohes Maß an regionaler Syntenieerhaltung zwischen dem *tebp-1*-Gen von *C. elegans* und anderen Spezies innerhalb der Gattung *Caenorhabditis*, während *tebp-2* keine Anzeichen regionaler Syntenie zeigte. Unsere Ergebnisse liefern erste Beweise für einen Komplex von Telomer-bindenden Proteinen in *C. elegans*, der die *C. elegans* Telomer-Doppelstrang-Binder R06A4.2 und T12E12.3 enthält.

4.10.3 Statement of Contribution








4.10.3 Statement of Contribution

As a co-author, I participated in the evolutionary analysis of *tebp-1* and *tebp-2*, which showed exclusive presence in the *Caenorhabditis* genus. My role in the study included extracting the sequences of the proteins from WormBase (WS275) and WormBase ParaSite (WBPS14/WS271), and conducting a BLASTP search to identify orthology relationships. To obtain a comprehensive view of their evolutionary history, I generated the multiple sequence alignment using MAFFT and performed the phylogenetic analysis using IQ-TREE for both the full alignment and the N-terminal region. The domain prediction of the RAP1 homeodomain was carried out using PFAM, and positive selection analysis was performed using DataMonkey. Additionally, I conducted the synteny analysis using genomic coordinate data. I critically read and commented on the manuscript.

Supervisor confirmation Falk Buitel

4.10.4 Article I: Main Text

The double-stranded DNA-binding proteins TEBP-1 and TEBP-2 form a telomeric complex with POT-1

Sabrina Dietz ^{1,6}, Miguel Vasconcelos Almeida ^{1,4,5,6}, Emily Nischwitz¹, Jan Schreier¹, Nikenza Viceconte ¹, Albert Fradera-Sola ¹, Christian Renz¹, Alejandro Ceron-Noriega¹, Helle D. Ulrich¹, Dennis Kappei ^{2,3}, René F. Ketting ¹ & Falk Butter ¹✉

Telomeres are bound by dedicated proteins, which protect them from DNA damage and regulate telomere length homeostasis. In the nematode *Caenorhabditis elegans*, a comprehensive understanding of the proteins interacting with the telomere sequence is lacking. Here, we harnessed a quantitative proteomics approach to identify TEBP-1 and TEBP-2, two paralogs expressed in the germline and embryogenesis that associate to telomeres in vitro and in vivo. *tebp-1* and *tebp-2* mutants display strikingly distinct phenotypes: *tebp-1* mutants have longer telomeres than wild-type animals, while *tebp-2* mutants display shorter telomeres and a Mortal Germline. Notably, *tebp-1;tebp-2* double mutant animals have synthetic sterility, with germlines showing signs of severe mitotic and meiotic arrest. Furthermore, we show that POT-1 forms a telomeric complex with TEBP-1 and TEBP-2, which bridges TEBP-1/-2 with POT-2/MRT-1. These results provide insights into the composition and organization of a telomeric protein complex in *C. elegans*.

¹Institute of Molecular Biology (IMB), Mainz, Germany. ²Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore. ³Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. ⁴Present address: Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, UK. ⁵Present address: Department of Genetics, University of Cambridge, Cambridge, UK. ⁶These authors contributed equally: Sabrina Dietz, Miguel Vasconcelos Almeida. ✉email: f.butter@imb-mainz.de

Most telomeres in linear eukaryotic chromosomes end in tandem repeat DNA sequences. Telomeres solve two major challenges of chromosome linearity: the end-protection problem and the end-replication problem^{1,2}. The end-protection problem originates from the structural similarity between telomeres and DNA double-strand breaks, which can lead to recognition of the telomere by the DNA damage surveillance machinery². When telomeres are falsely recognized as DNA damage, they are processed by the non-homologous end joining or homologous recombination pathways, leading to genome instability^{3,4}. The end-replication problem arises from the difficulties encountered by the DNA replication machinery to extend the extremities of linear chromosomes, which results in telomere shortening with every cell division^{5–7}. When a subset of telomeres shorten beyond a critical point, cellular senescence or apoptosis are triggered^{8–10}.

Specialized proteins have evolved to deal with the complications arising from telomeres, which in vertebrates are composed of double-stranded (ds) (TTAGGG)_n repeats ending in a single-stranded (ss) 3' overhang¹¹. In mammals, a telomere-interacting complex of six proteins termed shelterin constitutively binds to telomeres in mitotic cells¹². This complex consists of the ds telomere binders TRF1 and TRF2, the TRF2-interacting protein RAP1, the ss binding protein POT1 and its direct interactor TPP1, as well as the bridging protein TIN2. Altogether, the proteins of this complex shield telomeres from a DNA damage response by inhibiting aberrant DNA damage signaling³. In addition, shelterin components are required for the recruitment of the telomerase enzyme, which adds de novo repeats to the telomeric ends, allowing maintenance of telomere length in dividing cells⁶. Telomerase is a ribonucleoprotein, comprised of a catalytic reverse-transcriptase protein component and an RNA moiety. Besides the core shelterin complex, additional proteins have been described to interact with telomeres and assist in the maintenance of telomere length, e.g., HMBOX1 (also known as HOT1), ZBTB48 (also known as TZAP), NR2C2, and ZNF827^{13–17}.

In *Schizosaccharomyces pombe*, a shelterin-like complex harboring orthologs of the human shelterin complex was described^{18–20}. TAZ1 and POT1 bind to ds and ss telomeric DNA similar to their human counterparts TRF1/TRF2 and POT1, respectively. In turn, *Saccharomyces cerevisiae* has distinct complexes binding to the ds and ss telomere^{21–26}. The *S. cerevisiae* ortholog of the TRF2-interacting protein RAP1 binds ds telomeric DNA through two domains structurally related to Myb domains²⁷. The ss overhang is not bound by a POT1 homolog but rather by the CST complex^{22,23,25}. Overall, this indicates that different telomere-binding complexes have evolved across species to alleviate the challenges of linear chromosome ends, based on variations of recurring DNA-binding modules.

The nematode *Caenorhabditis elegans* has been employed in many seminal discoveries in molecular biology, genetics, and development²⁸. Its telomeres have a repeat sequence similar to vertebrate telomeres, consisting of (TTAGGC)_n²⁹. Moreover, *C. elegans* telomeres have a length of about 2–9 kb^{29,30}, and it has been proposed that its telomeric structures have both 5' and 3' ss overhangs, each recognized by dedicated ss telomere-binding proteins³¹. Telomere maintenance in this nematode is carried out by the catalytic subunit of telomerase TRT-1³². The RNA component of *C. elegans* telomerase has not been identified thus far. Telomeres can be maintained by additional mechanisms, since *C. elegans* can survive without a functioning telomerase pathway by employing alternative lengthening of telomere (ALT)-like mechanisms, creating more heterogeneous telomere lengths^{33–37}.

In *C. elegans*, four proteins with domains structurally similar to the DNA-binding domain of human POT1 were identified. Three

of those proteins, namely POT-1 (also known as CeOB2), POT-2 (also known as CeOB1), and MRT-1, were confirmed to bind to the ss telomeric overhangs^{31,38}. Mutants for these factors show telomere length maintenance defects. Depletion of POT-1 and POT-2 leads to telomere elongation^{31,33,35,37}, whereas depletion of MRT-1 results in progressive telomere shortening over several generations³⁸. Concomitant to telomere shortening, *mrt-1*, *mrt-2*, and *trt-1* mutant animals share a Mortal Germline (Mrt) phenotype, characterized by a gradual decrease in fertility across generations, until animals become sterile^{30,32,38}. MRT-1 was proposed to be in a pathway for facilitation of telomere elongation together with the DNA damage checkpoint protein MRT-2, and telomerase TRT-1³⁸. Despite the identification of these different telomere-associated proteins, no telomere-binding complex has been described in *C. elegans* yet.

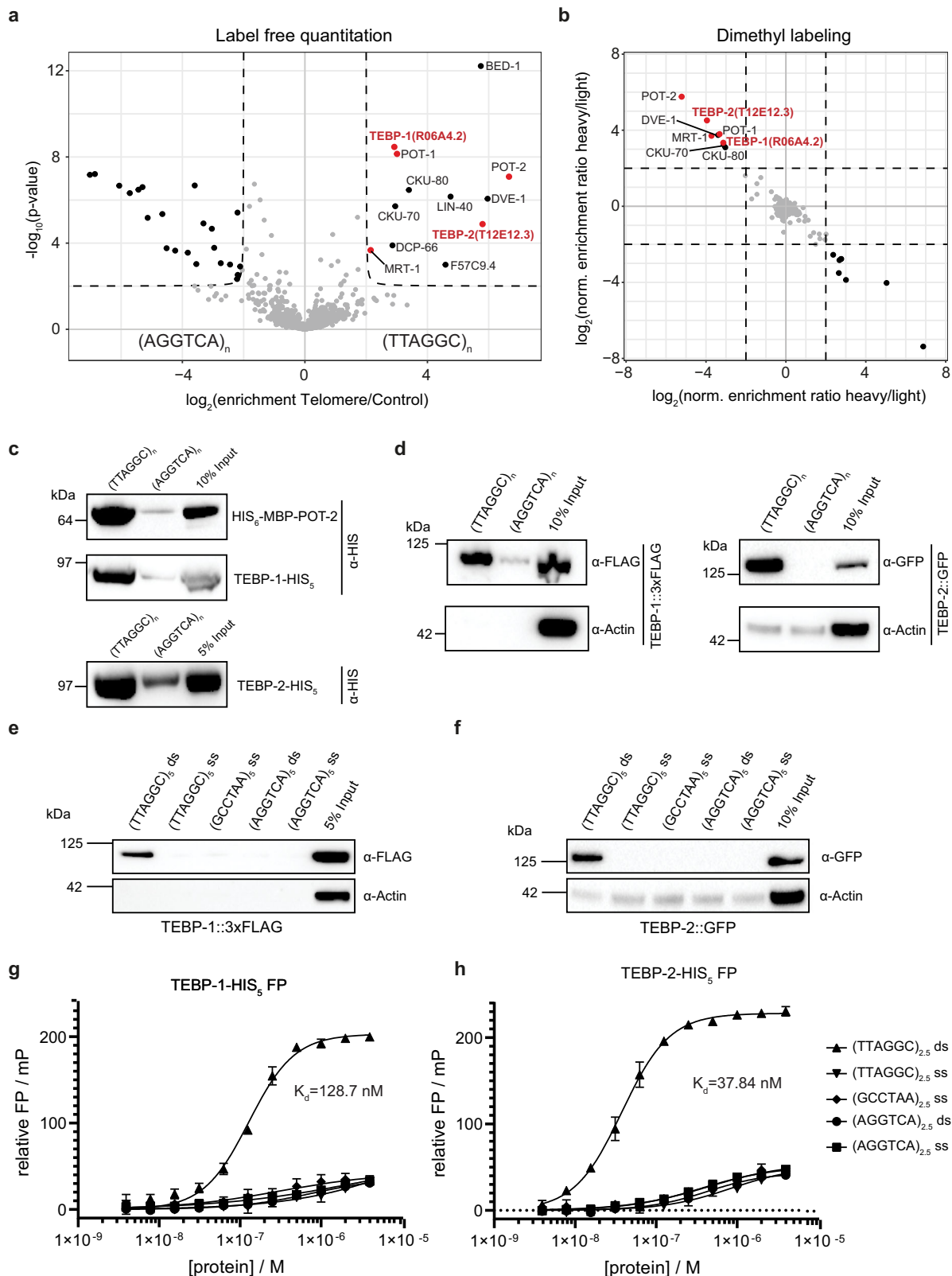
In this work, we performed a quantitative proteomics screen to identify novel telomere-binding proteins in *C. elegans*. We report the identification and characterization of R06A4.2 and T12E12.3, two previously uncharacterized paralog genes, which we named telomere-binding proteins 1 and 2 (*tebp-1* and *tebp-2*), respectively. TEBP-1 and TEBP-2 bind to the ds telomeric sequence in vitro with nanomolar affinity and co-localize with POT-1, a known telomere binder, in vivo. *tebp-1* and *tebp-2* mutants have contrasting effects on telomere length: while *tebp-1* mutants display elongated telomeres, *tebp-2* mutants have shortened telomeres. In addition, TEBP-1 and TEBP-2 have important roles in fertility, as *tebp-1*; *tebp-2* double mutants are synthetic sterile. Size-exclusion chromatography and interaction studies demonstrate that TEBP-1 and TEBP-2 are part of a complex with POT-1, which bridges the ds telomere binders, TEBP-1 and TEBP-2, with the ss telomere binders POT-2 and MRT-1.

Results

TEBP-1 (R06A4.2) and TEBP-2 (T12E12.3) are double-stranded telomere-binding proteins in *Caenorhabditis elegans*. To identify proteins that bind to the *C. elegans* telomeric sequence, we employed a DNA pulldown assay (Supplementary Fig. 1a, b) previously used to successfully identify telomeric proteins in other species^{15,16,39,40}. We incubated concatenated, biotinylated DNA oligonucleotides consisting of either the telomeric sequence of *C. elegans* (TTAGGC)_n, or a control sequence (AGGTCA)_n, with nuclear-enriched extracts of gravid adult worms. The experiment was performed twice using two different quantitative proteomics approaches: label-free quantitation (LFQ)⁴¹ and reductive dimethyl labeling (DML)⁴², which yielded 12 and 8 proteins enriched in telomeric sequence pulldowns, respectively, with an overlap of 8 proteins (Fig. 1a, b and Supplementary Fig. 1a, b). Among these eight proteins, we found the already known ss telomere binders POT-1, POT-2, and MRT-1^{31,33,37,38}, as well as the CKU-70/CKU-80 heterodimer⁴³, and three additional proteins: R06A4.2, T12E12.3, and DVE-1.

R06A4.2 and T12E12.3 were of particular interest, as they share 74.3% DNA coding sequence identity and 65.4% amino acid sequence identity (Supplementary Fig. 1c), suggesting that R06A4.2 and T12E12.3 are paralogs. While R06A4.2 and T12E12.3 lack any annotated protein domain, using HHpred v3.2.0⁴⁴, we could determine that the N-terminal region of both proteins shows similarity to the homeodomains of human and yeast RAP1 (Supplementary Fig. 1d, e and Supplementary Data file 1). RAP1 is a direct ds telomere binder in budding yeast^{21,45}, and a member of the mammalian shelterin complex through interaction with TRF2⁴⁶.

We validated binding of R06A4.2 and T12E12.3 to telomeric DNA by performing DNA pulldowns with His-tagged recombinant proteins (Fig. 1c). Using CRISPR-Cas9 genome editing, we



inserted a *gfp* and a *3xflag* sequence directly upstream of the endogenous stop codon of *T12E12.3* and *R06A4.2*, respectively (Supplementary Fig. 1d, e). Using these strains, we could show that the endogenously tagged versions of *R06A4.2* and *T12E12.3* also bind to the *C. elegans* telomere sequence (Fig. 1d).

Owing to the preparation strategy, our concatenated DNA probes contained both ds and ss DNA, which precludes any

conclusions about whether *R06A4.2* and *T12E12.3* bind ss or ds telomeric DNA. We thus performed additional DNA pull-downs with ss and ds probes specifically designed with five repeats (TTAGGC)₅. Both proteins were found to exclusively bind to the ds telomeric repeats, establishing *R06A4.2* and *T12E12.3* as ds telomeric binders (Fig. 1e, f). To confirm and quantify the interaction of *R06A4.2* and *T12E12.3* with ds telomeric DNA, we

Fig. 1 **TEBP-1 (R06A4.2) and TEBP-2 (T12E12.3) are double-stranded telomere binders in *C. elegans*.** **a** Volcano plot representing label-free proteomic quantitation of pull-downs with biotinylated, concatenated oligonucleotide baits of telomeric DNA sequence (TTAGGC)_n or control DNA sequence (AGGTCA)_n. Pull-downs were performed with nuclear extracts from synchronized gravid adult animals, in octuplicates per condition (two biological replicates, each with four technical replicates). Log₂ fold enrichment of proteins in one condition over the other is presented on the x-axis. The y-axis shows $-\log_{10} p$ -value (Welch *t*-test) of enrichment across replicates. More than 4-fold enriched proteins with *p*-value < 0.01 are annotated as black dots, the background proteins as gray dots. Enriched proteins of interest, such as the known ss telomere binders, are annotated as red dots. **b** Scatterplot representing results of reductive dimethyl-labeling-based quantitation of pull-downs with the same extract and DNA baits as in (a). Per condition, pull-downs were performed in duplicates and labeled on the peptide level, including an intra-experimental label switch to achieve cross-over sets. The x-axis represents log₂ transformed ratios of the reverse experiment, whereas the y-axis represents log₂ transformed ratios of the forward experiment (see Supplementary Fig. 1b). Single proteins are depicted by dots in the scatterplot. Enriched proteins (threshold > 4) are annotated as black dots, background proteins as gray dots, and enriched proteins of interest as red dots. **c** Binding of recombinant His-tagged POT-2, TEBP-1 and TEBP-2, from crude *E. coli* lysate, to telomere or control DNA as in (a). Chemiluminescence western blot read-out, after probing with α -His antibody. POT-2 is used as a positive control for telomeric repeat binding. MBP: Maltose-binding protein, kDa: kilodalton. Uncropped blots in Source Data. *N* = 2 biologically independent experiments with similar results, except POT-2 *N* = 1. **d** DNA pull-downs as in c but on embryo extracts of transgenic *C. elegans* lines carrying either TEBP-1::3xFLAG or TEBP-2::GFP. *N* = 2 independent experiments with similar results, **e, f** DNA pull-downs with 5x telomeric (TTAGGC) double-strand (ds) repeats and both respective single-strand (ss) baits, and 5x control (AGGTCA) ds or 5x (AGGTCA) ss repeats. Pull-downs were performed with embryo extracts of TEBP-1::3xFLAG or TEBP-2::GFP animals. Uncropped blots in Source Data. *N* = 3 biologically independent experiments with similar results, **g, h** Fluorescence polarization assays of 4 μ M to 4 nM purified TEBP-1-His₅ and TEBP-2-His₅, respectively. Binding affinities to 2.5x ss and ds telomeric and control repeats of FITC-labeled oligonucleotides. Error bars represent \pm the standard deviation of the mean values. Per data point *n* = 3 technical replicates. FP, fluorescence polarization; mP, millipolarization, upward triangle: 2.5x TTAGGC double-strand, downward triangle: 2.5x TTAGGC single-strand, diamond: 2.5x GCCTAA single-strand, circle: 2.5x shuffled control double-strand, square: 2.5x shuffled control single-strand.

performed fluorescence polarization with purified, recombinant proteins and FITC-labeled oligonucleotides. Both T12E12.3 and R06A4.2 displayed affinity for the ds telomeric repeat sequence in the nanomolar range ($K_d = 128.7$ nM for R06A4.2 and $K_d = 37.84$ nM for T12E12.3, Fig. 1g, h). Both T12E12.3 and R06A4.2 showed highest affinity for the 2.5x telomeric repeat, when incubated with a 2.5x, 2.0x, 1.5x T-rich, and 1.5x G-rich telomeric repeat sequences (Supplementary Fig. S2a–c).

In conclusion, we demonstrate that R06A4.2 and T12E12.3, two proteins with highly similar sequence, bind directly and with high affinity to the *C. elegans* ds telomeric DNA sequence in vitro. Thus, we decided to name R06A4.2 as Telomere-Binding Protein-1 (TEBP-1) and T12E12.3 as Telomere-Binding Protein-2 (TEBP-2).

TEBP-1 and TEBP-2 localize to telomeres in proliferating cells in vivo. To explore the expression pattern of *tebp-1* and *tebp-2* throughout animal development, we used a recently published mRNA-seq dataset⁴⁷. Both genes show the highest expression in embryos, very low abundance during the L1–L3 larval stages, and an increase in expression in L4 larvae and young adults (YAs, Supplementary Fig. 3a–c). The observed increase in *tebp-1* and *tebp-2* mRNA expression from the L4 to YA stages coincides with the increased progression of germline development, which may hint to a higher expression level during gametogenesis. Indeed, using available gonad-specific RNA-seq datasets⁴⁸, we confirmed that *tebp-1* and *tebp-2* are expressed in spermatogenic and oogenic gonads (Supplementary Fig. 3d). Similar developmental mRNA expression patterns were also found for the known ss telomere binders *pot-1*, *pot-2*, and *mrt-1* (Supplementary Fig. 3a, d). To study the expression at the protein level, we crossed our endogenously tagged strains to generate a *tebp-1::3xflag*; *tebp-2::gfp* strain to monitor protein abundance simultaneously by western blot. The protein expression patterns of TEBP-1 and TEBP-2 are highly similar to the RNA-seq data, with highest detected expression in embryos, a drop during the larval stages L1–L4, ultimately followed by an increase in YA (Fig. 2a).

To study TEBP-1 and TEBP-2 localization in vivo, we focused on embryos and on the germline of adult animals. In these two actively dividing tissues, TEBP-1 and TEBP-2 protein expression is high and condensed chromosomes facilitate visualization of telomeric co-localization. In addition to the *tebp-2::gfp* strain used above, we also generated an endogenously tagged *tebp-1::gfp*

allele, using CRISPR-Cas9 genome editing (Supplementary Fig. 1d). To check for telomeric localization in vivo, we crossed *tebp-1::gfp* and *tebp-2::gfp* each with a germline-specific *pot-1::mCherry* single-copy transgene³⁷, and imaged the dual-fluorescent animals. TEBP-1::GFP and TEBP-2::GFP co-localize with POT-1::mCherry inside the nuclei of oocytes and embryos (Fig. 2b–e). Confocal microscopy of TEBP-1::GFP in combination with POT-1::mCherry was challenging likely due to bleaching of TEBP-1::GFP. Co-localization of TEBP-2::GFP and POT-1::mCherry was also observed in the mitotic region of the germline and in mature sperm (Fig. 2d). These results clearly establish that TEBP-1 and TEBP-2 co-localize with a known telomeric binder in vivo in proliferating tissues, indicating that their ability to bind ds telomeric DNA in vitro may have functional relevance.

TEBP-1 and TEBP-2 have opposing telomere length phenotypes. As TEBP-1 and TEBP-2 localize to telomeres, we sought to address whether these proteins regulate telomere length, as is the case for the known ss telomere-binding proteins POT-1, POT-2, and MRT-1^{31,33,37,38}. Using CRISPR-Cas9 genome editing, we generated *tebp-1* and *tebp-2* deletion mutants encoding truncated transcripts with premature stop codons (Supplementary Fig. 1d–g and Supplementary Fig. 4a, b). *tebp-1* and *tebp-2* mutants are viable and show no immediate, obvious morphological or behavioral defects. We analyzed telomere length in the mutants after propagation for more than 100 generations, sufficient to establish a “steady-state” telomere length phenotype, by carrying out a telomere Southern blot on mixed-stage animals. Interestingly, while *tebp-1(xf133)* shows an elongated telomere phenotype comparable to the *pot-2(tm1400)* mutant, *tebp-2(xf131)* shows a shortened telomere phenotype (Fig. 3a), similar to *mrt-1* mutants³⁸. In addition, we performed quantitative fluorescence in situ hybridization (qFISH) in dissected adult germlines, which confirmed our initial observation that *tebp-1* and *tebp-2* mutants display longer or shorter telomeres than wild-type, respectively (Fig. 3b–f). Furthermore, we also measured telomere length in embryos by qFISH. Like in the germline, the telomeres of *tebp-1* mutant embryos are elongated, while the telomeres of *tebp-2* embryos are shortened (Supplementary Fig. 4c–g).

In summary, *tebp-1* and *tebp-2* mutants display opposing regulatory effects on telomere length. These experiments suggest that the TEBP-1 protein counteracts telomere elongation

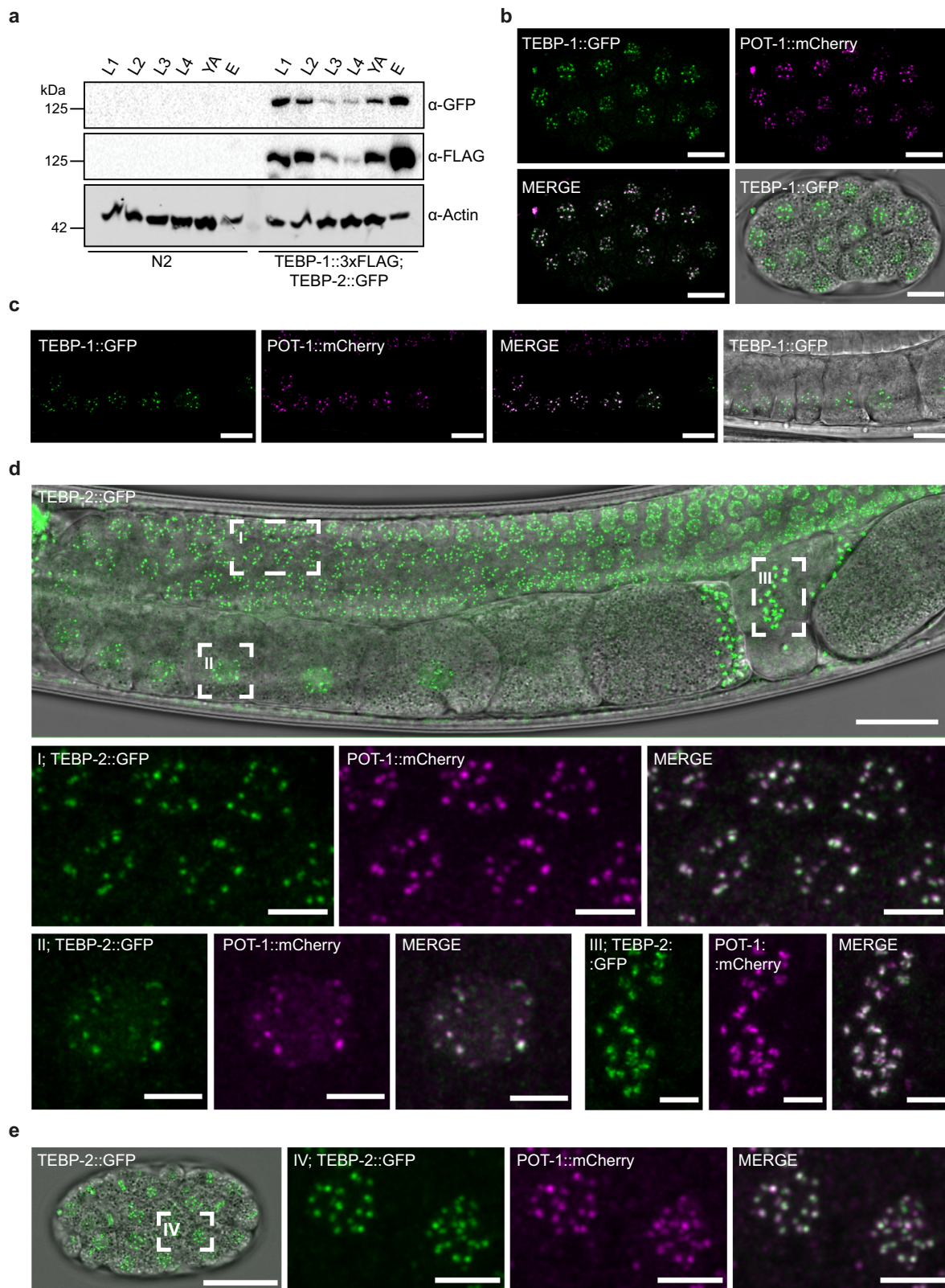


Fig. 2 TEBP-1 and TEBP-2 are expressed throughout *C. elegans* development and localize to telomeres in vivo. **a** Western blot of TEBP-1::3xFLAG and TEBP-2::GFP expression in different developmental stages of *C. elegans*. Thirty-five micrograms of extract from either N2 or a double transgenic line carrying TEBP-1::3xFLAG and TEBP-2::GFP were used. Actin was used as loading control. kDa: kilodalton. Uncropped blot in Source Data. $N = 1$ **b, c** Maximum intensity projections of representative confocal z-stacks of an embryo (**b**), or oocytes (**c**) expressing endogenously tagged TEBP-1::GFP and transgenic POT-1::mCherry. Scale bars, 10 μm . **d, e** Maximum intensity projections of representative confocal z-stacks of an adult animal (**d**), or embryo (**e**) expressing both endogenously tagged TEBP-2::GFP and transgenic POT-1::mCherry. Insets show nuclear co-localization in meiotic germ cell nuclei (I), an oocyte (II), spermatozoa (III), and embryonic cells (IV). Scale bars, 20 μm (overview) and 4 μm (insets). All microscopy images were deconvoluted using Huygens remote manager. Representative images from two individual animals per strain, $N = 2$ biologically independent experiments with similar results.

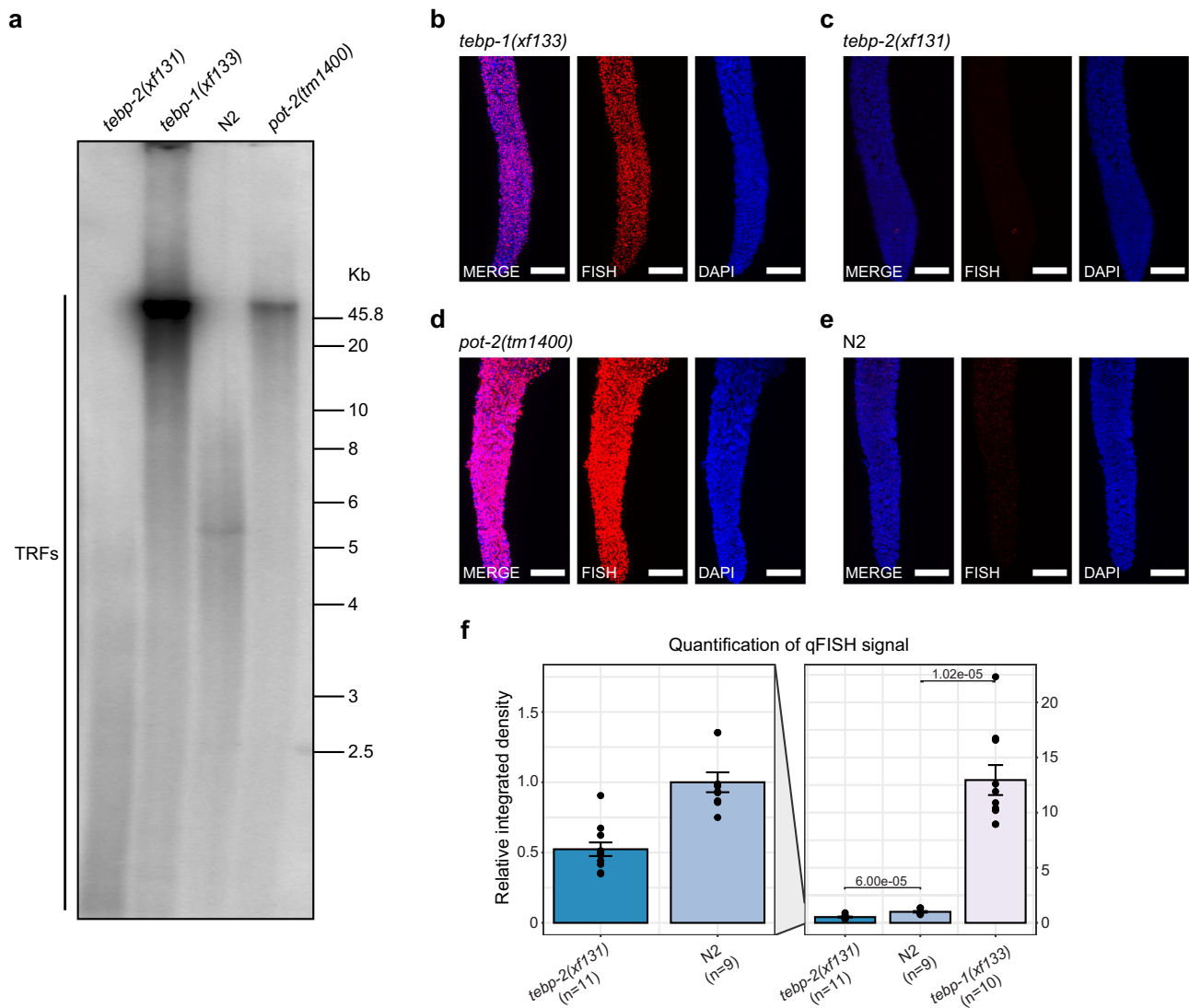
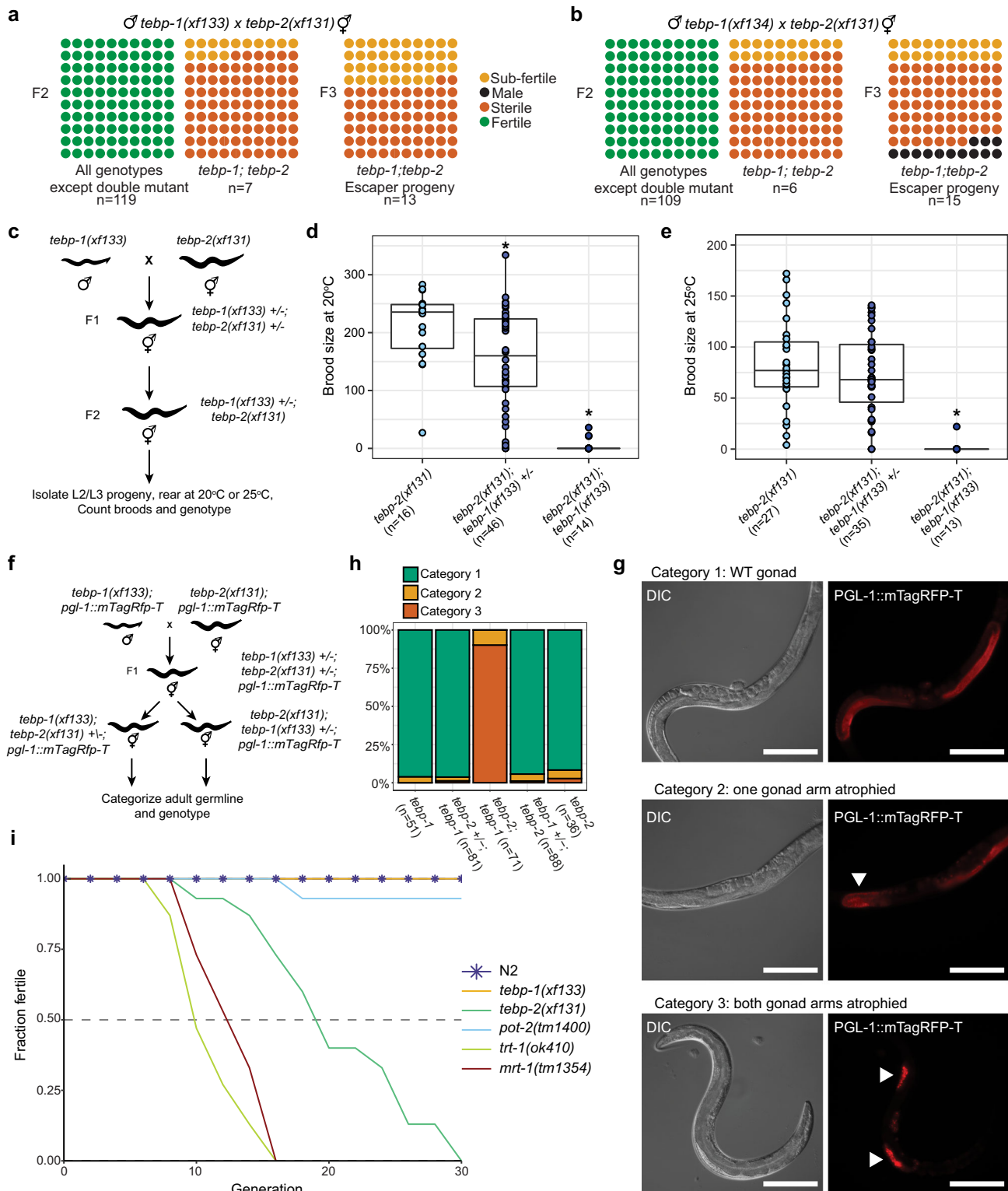


Fig. 3 TEBP-1 and TEBP-2 regulate telomere length. **a** Southern blot analysis of *C. elegans* telomeres. DNA from four different strains (*tebp-1(xf133)* grown for -102 generations; *tebp-2(xf131)*, grown for -124 generations; N2, and *pot-2(tm1400)*) was digested and separated by agarose gel electrophoresis. DNA was transferred to a positively charged nylon membrane and hybridized with a radiolabeled (GCCTAA)₃ oligonucleotide. Brightness and contrast of the membrane read-out were adjusted using Fiji. Telomere restriction fragments (TRFs) are indicated in the Fig.. Uncropped blot in Source Data. *N* = 3 independent experiments with similar results. **b–e** Representative maximum projection z-stacks of a qFISH assay using dissected adult germlines of the following *C. elegans* mutant strains: *tebp-1(xf133)* (grown for -98 generations), *tebp-2(xf131)* (grown for -120 generations), *pot-2(tm1400)*, and wild-type N2. The telomeres of dissected worms of the respective strains were visualized by hybridization with a telomeric PNA-FISH-probe. Nuclei were stained with DAPI. Scale bars, 15 μ m. **f** Barplot depicting analysis of qFISH images of the strains in (**b–c**) and (**e**). Average telomere length is indicated by arbitrary units of relative integrated density, with wild-type N2 set to 1. The plot on the left shows the *tebp-2(xf131)* and N2 values zoomed-in. Analyzed *n* per strain derived from independent animals: *tebp-2(xf131)*: *n* = 11, N2: *n* = 9, *tebp-1(xf133)*: *n* = 10. Error bars represent the standard error of the mean (SEM) and *p*-values were calculated using Welch's *t*-test. *N* = 3 biologically independent experiments with similar results.

independently of telomerase, while TEBP-2 promotes telomere lengthening.

Simultaneous lack of TEBP-1 and TEBP-2 leads to synthetic sterility. To better understand how *tebp-1* and *tebp-2* mutants distinctly affect telomere length, we intended to measure telomere length in *tebp-1*; *tebp-2* double mutants. Surprisingly, when we crossed our single mutants, we could not establish a double homozygous *tebp-1*; *tebp-2* mutant strain. In fact, *tebp-1*; *tebp-2* double mutants displayed highly penetrant synthetic sterility (Fig. 4a). Repeating the cross with another *tebp-1* mutant allele (*xf134*), as well as the reciprocal cross, yielded the same synthetic sterility (Fig. 4b and Supplementary Fig. 5a). Only about 14–38%

of F2 or F3 *tebp-1*; *tebp-2* animals did not have synthetic sterility (Fig. 4a, b). These “synthetic sterility escapers” were subfertile, siring less than 60 offspring. Importantly, a *tebp-2::gfp* single-copy transgene fully rescued the appearance of sterility, demonstrating that the C-terminal tag does not disrupt TEBP-2 function (Supplementary Fig. 5a). When we combined *tebp-1* mutant animals with *mrt-1*, *trt-1*, or *pot-2* mutations, or *tebp-2* mutant animals with *trt-1* or *pot-2*, the double mutant offspring was fertile (Supplementary Fig. 5a). These results demonstrate that the synthetic sterility is specific to *tebp-1*; *tebp-2* double mutants, and is not a consequence of crossing shorter telomere mutants with longer telomere mutants. We further quantified the synthetic sterility on brood size by picking L2–L3 progeny of *tebp-2*; *tebp-1* +/- mutants, blind to genotype and germline health, rearing those



animals at 20 °C or 25 °C, later counting their brood sizes, and genotyping each animal (Fig. 4c–e). This revealed that the immediate synthetic sterility phenotype is not dependent on temperature, as the reduction of progeny numbers was apparent at both 20 and 25 °C.

Morphologically, *tebp-1; tebp-2* double mutants displayed a degenerated germline. To visualize this phenotype, we created *tebp-1* and *tebp-2* strains in combination with an endogenously tagged *pgl-1::mTagRfp-T* allele^{49,50}, which we used as a germ

cell reporter. PGL-1 is expressed in P-granules, perinuclear granules most important for germline development and gene regulation^{51,52}. As depicted in Fig. 4f, we repeated the *tebp-1* x *tebp-2* cross with *pgl-1::mTagRfp-T* in the background, isolated cross progeny of the indicated genotypes, reared these animals to adulthood, scored them into three categories of germline morphology, and genotyped them afterwards. The categories can be described as follows: category 1 animals displayed a wild-type or near wild-type morphology (Fig. 4g, upper panels),

Fig. 4 *tebp-1*; *tebp-2* double mutants have synthetic sterility, and *tebp-2* mutants have a Mortal Germline. **a, b** Schematics depicting the quantification of fertility of the F2 (two panels on the left) and F3 (panel on the right) cross progeny of the indicated crosses. Each dot represents 1% of the indicated *n* per square, in a 10 × 10 matrix for 100%. Green dots indicate fertile worms, yellow dots subfertile worms (<60 progeny), orange dots sterile worms, and black dots indicate male worms. The F3 animals used for the panels on the right were the progeny of subfertile F2s, which escaped synthetic sterility. Males with two different *tebp-1* mutant alleles, *xf133* and *xf134*, were used in **(a)** and **(b)**, respectively. **c** Schematic of cross performed with *tebp-1(xf133)* and *tebp-2(xf131)* to isolate progeny for determination of brood size at 20 and 25 °C. **d, e** Brood sizes of cross progeny animals, isolated as indicated in **(c)**, which were grown at 20 °C **(d)**, or 25 °C **(e)**. Central horizontal lines represent the median, the bottom and top of the box represent the 25th and 75th percentile, respectively. Whiskers represent the 5th and 95th percentile, dots represent the data points used to calculate the box plot. *n* is indicated on the x-axis label. In **(d)**, asterisks indicate the *p*-values of 9.6e-03 and 2.5e-06, as assessed by two-sided, unpaired Mann-Whitney and Wilcoxon tests comparing *tebp-1* worms with the cross siblings of the other genotypes. In **(e)**, asterisk indicates *p*-value = 4.1e-07, computed as in **(d)**. **f** Schematic of a repetition of the double mutant cross as in **(c)** with *pgl-1::mTagRfp-T* in the background. Worms heterozygous for one of the *tebp* mutations were singled and their germline categorized at day 2-3 of adulthood, according to germline morphology and assessed by PGL-1::mTagRFP-T expression. Worms were genotyped after categorization and imaging. **g** Representative widefield differential interference contrast (DIC) and fluorescence pictures of the three germline morphology categories defined. Scale bars, 200 μm. Atrophied germlines in categories 2 and 3 are marked with a white arrowhead. **h** Barplot representing the quantification of each category, per genotype as indicated on the x-axis. Number of animals analyzed is shown in the x-axis labels. **i** Plot showing the fraction of fertile populations of each indicated genotype across successive generations grown at 25 °C. *n* = 15 populations per strain.

category 2 animals displayed one atrophied gonad arm (Fig. 4g, middle panels), and category 3 animals had both gonad arms atrophied (Fig. 4g, lower panels). Besides Fig. 4g, representative animals for categories 2 and 3 are shown in Supplementary Fig. 5b. More than 85% of *tebp-1*; *tebp-2*; *pgl-1::mTagRfp-T* worms had a category 3 germline, while the remainder had only one gonad arm atrophied (Fig. 4h). Atrophied gonads generally showed under-proliferation of the germ cell nuclei of the mitotic zone and rare entry into meiosis, suggesting severe defects in cell division (Fig. 4g and Supplementary Fig. 5b). In addition, almost 15% (17/114 animals) of the progeny of *tebp-1*; *tebp-2*; *pgl-1::mTagRfp-T* synthetic sterility escapers were males, indicative of a high incidence of males (Him) phenotype. The synthetic sterility escaper progenies of previous crosses were also Him, at least in some cases (see F3 escaper progeny in Fig. 4b). Lastly, approximately 8% (8/97) of hermaphrodite *tebp-1*; *tebp-2*; *pgl-1::mTagRfp-T* escaper progeny had growth defects: while some reached adulthood but remained smaller than wild-type, others arrested prior to adulthood (Supplementary Fig. 5c).

Overall, these data show that the lack of functional TEBP-1 and TEBP-2 leads to severe germline defects that impede germline development.

TEBP-2 is required for transgenerational fertility. Despite the synthetic sterility of the double mutants, *tebp-1* and *tebp-2* single mutants did not have a baseline reduction in fertility when grown at 20 and 25 °C (Supplementary Fig. 5d, e). Nevertheless, mutants of telomere regulators, like *trt-1* and *mrt-1*, exhibit a Mrt phenotype, characterized by progressive loss of fertility across many generations^{32,38}. We thus conducted a Mortal Germline assay at 25 °C using late generation mutants, and found that *tebp-1* and *tebp-2* mutants displayed opposing phenotypes in line with their differing effects on telomere length. While *tebp-1(xf133)* remained fertile across generations, like wild-type, *tebp-2(xf131)* showed a Mrt phenotype (Fig. 4i), the onset of which is delayed compared to *mrt-1(tm1354)* and *trt-1(ok410)*, indicating a slower deterioration of germline health over generations. These results show that TEBP-2 is required to maintain germline homeostasis transgenerationally, while TEBP-1 is not.

TEBP-1 and TEBP-2 are part of a telomeric complex in *C. elegans*. Our initial mass spectrometry approach allowed us to identify proteins associated with the telomeres of *C. elegans*. However, it remains unknown if these factors interact and whether they are part of a telomere-binding complex. To address this,

we performed size-exclusion chromatography with embryonic extracts from a strain expressing TEBP-1::3xFLAG; TEBP-2::GFP. Western blot analysis of the eluted fractions shows that TEBP-1 and TEBP-2 have very similar elution patterns with one peak ranging from 450 kDa to 1.5 MDa, with a maximum at 1.1 MDa (Fig. 5a and Supplementary Fig. 6a). Next, we reasoned that the elution peak would shift if telomeric DNA is enzymatically degraded. To test this, embryonic extracts were treated with *Serratia marcescens* nuclease (Sm nuclease), a non-sequence-specific nuclease, prior to size-exclusion chromatography, but we did not observe a strong shift (Fig. 5b). While we cannot fully exclude the possibility that telomeric DNA was inaccessible to Sm nuclease digestion, the results suggest that TEBP-1 and TEBP-2 are part of a telomeric complex.

To identify proteins interacting with TEBP-1 and TEBP-2, we performed immunoprecipitation (IP) followed by quantitative mass spectrometry (qMS) in embryos (Fig. 5c, d) and YAs (Supplementary Fig. 6b, c). Notably, IP-qMS of TEBP-1 and TEBP-2 baits enriched for MRT-1, POT-1, and POT-2, the three known ss telomere-binding proteins in *C. elegans*. In some cases, (Fig. 5d and Supplementary Fig. 6b) it was difficult to unambiguously assign unique peptides to TEBP-1::3xFLAG and TEBP-2::GFP in our qMS analysis, given their high protein sequence identity (65.4%). However, we confirmed by co-IP experiments that TEBP-1 and TEBP-2 reciprocally interact in embryos and YA (Fig. 5e, f and Supplementary Fig. 6d). Moreover, TEBP-1 and TEBP-2 remain associated with MRT-1, POT-1, and POT-2 even after treatment with Sm nuclease (Supplementary Fig. 6e, f).

POT-1 is required to bridge the double-stranded and the single-stranded telomere. To reveal the architecture of the telomeric complex, we sought to identify direct interactions amongst TEBP-1, TEBP-2, POT-1, POT-2, and MRT-1, using a yeast two-hybrid (Y2H) screen. While TEBP-2 fused to the DNA-binding domain of Gal4 unfortunately self-activated the reporter (Supplementary Fig. 6g), we could identify direct interactions of POT-1 with TEBP-1 and TEBP-2 (Fig. 6a and Supplementary Fig. 6g). Furthermore, in accordance with IP-qMS and co-IP experiments (Fig. 5e, f and Supplementary Fig. 6d), we confirmed interaction between TEBP-1 and TEBP-2 in the Y2H experiment (Fig. 6a and Supplementary Fig. 6g). These results are consistent with a scenario where TEBP-1 and TEBP-2 interact directly with each other and with POT-1.

The observed direct interactions suggest that POT-1 may be a critical link between the ds and the ss telomeric region. To test this idea, we performed IP-qMS of TEBP-1 and TEBP-2, in

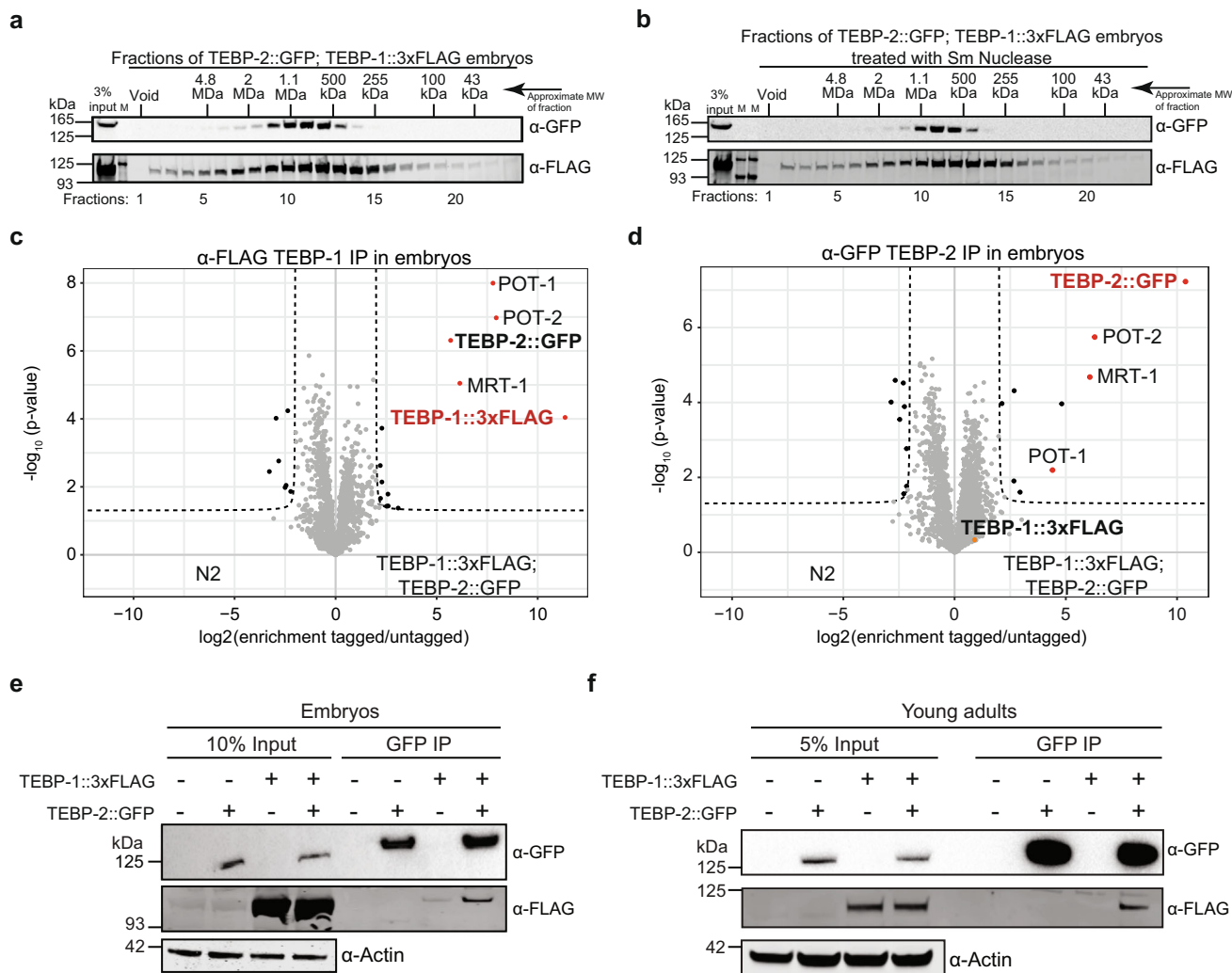


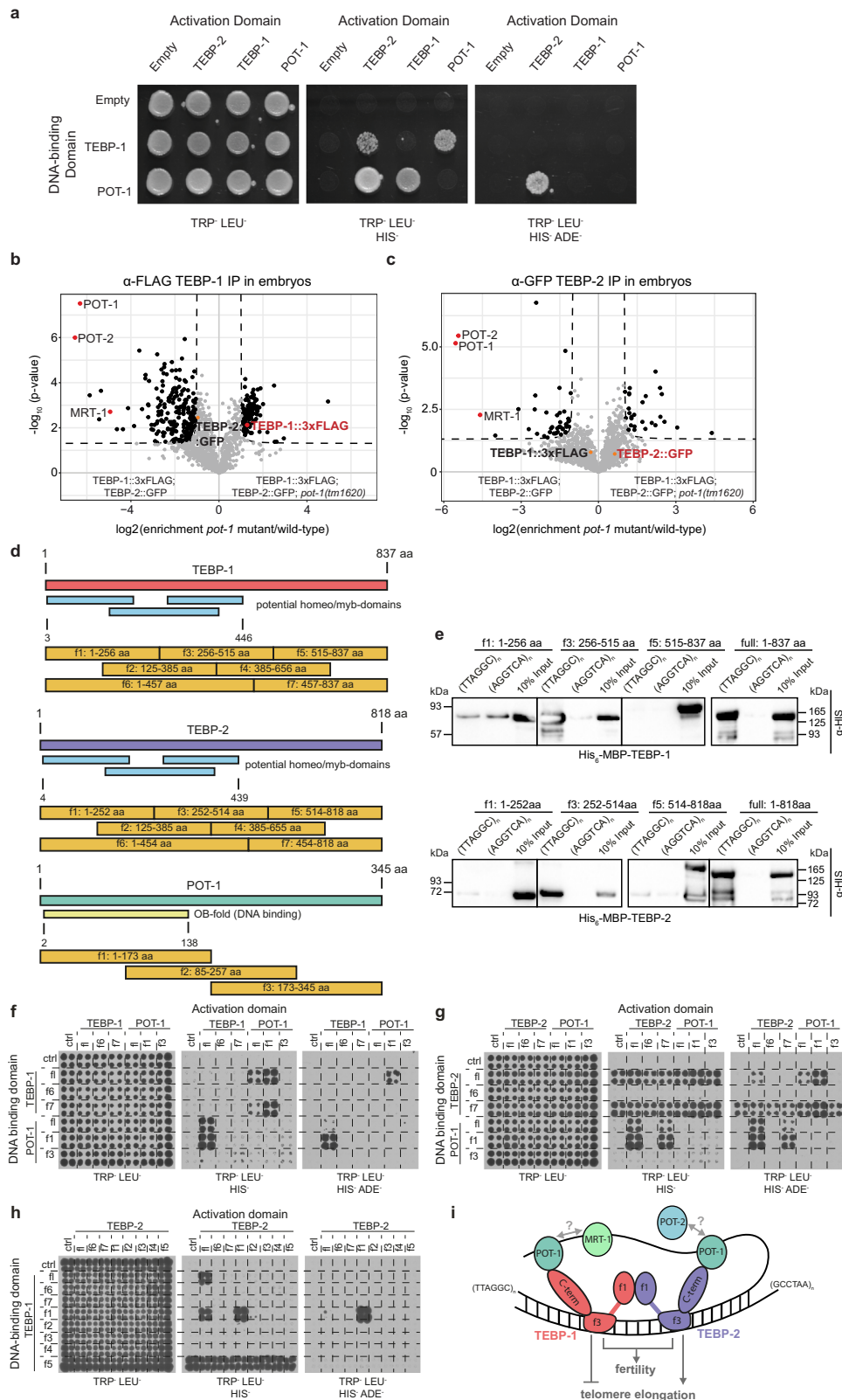
Fig. 5 TEBP-1 and TEBP-2 are part of a telomeric protein complex. **a** Size-exclusion chromatography of embryo extracts expressing TEBP-1::3xFLAG and TEBP-2::GFP, followed by western blot of the eluted fractions. The approximate molecular weight (MW) of the fractions is indicated on the Fig. panel. $N = 2$ biologically independent experiments with similar results. **b** Identical to **(a)**, but with treatment of embryo extracts with Sm nuclease, prior to size-exclusion chromatography. $N = 1$. **c, d** Volcano plots showing quantitative proteomic analysis of either TEBP-1::3xFLAG **(c)** or TEBP-2::GFP **(d)** IPs in embryos. IPs were performed in quadruplicates. Enriched proteins (threshold: 4-fold, $p\text{-value} < 0.05$) are shown as black dots, enriched proteins of interest are highlighted with red or orange dots, and the baits are named in red. Background proteins are depicted as gray dots. **e** Co-IP western blot experiment of TEBP-1::3xFLAG and TEBP-2::GFP. The IP was performed with a GFP-trap, on embryo extracts from strains carrying either one or both of the endogenous tags and wild-type. Actin was used as loading control. **f** Same co-IP experiment as in **(e)** but carried out with extracts from young adult worms. For **(e)** and **(f)** $N = 3$ biologically independent experiments with similar results.

wild-type and mutant *pot-1* backgrounds. These experiments showed that interaction of the ds telomere binders TEBP-1 and TEBP-2 with the ss binders POT-2 and MRT-1, is strongly depleted in *pot-1* mutants (Fig. 6b, c). TEBP-1 and TEBP-2 protein levels are not affected by the *pot-1* mutation, indicating the loss of interaction with POT-2 and MRT-1 is not due to reduced availability of TEBP-1 or TEBP-2 (Supplementary Fig. 6h). In addition, TEBP-1 and TEBP-2 still interact with each other in the absence of POT-1 (Supplementary Fig. 6h).

Next, to map the amino acid sequences responsible for TEBP-1 and TEBP-2 DNA-binding and protein-protein interactions, with each other and with POT-1, we divided their protein sequences into seven fragments (f1–f7), and the protein sequence of POT-1 into three fragments (f1–f3, Fig. 6d). DNA pull-downs with His-MBP-tagged TEBP-1 and TEBP-2 recombinant proteins demonstrated DNA binding by their f3 fragments (Fig. 6d, e), which contain their third predicted homeo-/myb-domain. Furthermore, Y2H experiments using the fragments shown in Fig. 6d, indicate

that the C-terminal tails of TEBP-1 and TEBP-2 (f7) interact with the OB-fold of POT-1 (Fig. 6f, g). Additional Y2H assays demonstrate that TEBP-1 and TEBP-2 interact with each other via their respective f1 fragments, encompassing their first predicted homeo-/myb-domains (Fig. 6h and Supplementary Fig. 6i).

Altogether, our data strongly indicate that TEBP-1 and TEBP-2 are integral parts of a telomeric complex, or complexes, which also include the known ss telomere binders POT-1, POT-2, and MRT-1. We propose a simple working model where TEBP-1 and TEBP-2 bind to the ds telomere via their third predicted homeo-/myb-domains, have opposed effects on telomere dynamics, and are required for fertility (Fig. 6i). POT-1, with the ability of its OB-fold to directly bind the C-terminal tails of TEBP-1 and TEBP-2 (Fig. 6a, f, g), as well as ss telomeric repeats *in vitro*³¹, may link the ds binders to the ss telomere, thereby bringing TEBP-1 and TEBP-2 in close proximity of POT-2 and MRT-1 (Fig. 6i).



Conservation of *tebp* genes in the *Caenorhabditis* genus. To infer the evolutionary history of *tebp-1* and *tebp-2* genes, we identified protein-coding orthologs by reciprocal BLASTP analysis in the searchable genomes in Wormbase and Wormbase ParaSite databases. Then, we performed a multiple sequence alignment with the ortholog protein sequences, and used it to build a phylogenetic tree (Fig. 7a and Supplementary Data file 2).

Our findings suggest that *tebp* orthologs are present only in the *Caenorhabditis* genus, mostly in the *Elegans* supergroup (which includes the *Elegans* and *Japonica* groups). A distinct number of protein-coding *tebp* genes was identified per species: *C. briggsae*, *C. nigoni*, *C. sinica*, and *C. japonica* have one *tebp* ortholog; *C. elegans*, *C. inopinata*, *C. remanei*, *C. brenneri*, *C. tropicalis*, and *C. angaria* have two *tebp* orthologs; and *C. latens* has three *tebp*

Fig. 6 POT-1 links the ds telomere binders to the ss telomere. **a** Y2H assay with full length TEBP-1, TEBP-2, and POT-1 fusions to the activation or DNA-binding domains of Gal4. Growth on TRP⁻ LEU⁻ HIS⁻ plates demonstrates interaction. Growth on high stringency TRP⁻ LEU⁻ HIS⁻ ADE⁻ medium suggests strong interaction. TRP⁻: lacking tryptophan, LEU⁻: lacking leucine, HIS⁻: lacking histidine, ADE⁻: lacking adenine. **b, c** Volcano plots showing quantitative proteomic analysis of either TEBP-1::3xFLAG (**b**) or TEBP-2::GFP (**c**) IPs in embryos. IPs were performed in quadruplicates. Enriched proteins (threshold: 2-fold, p -value < 0.05) are shown as black dots, enriched proteins of interest are highlighted with red or orange dots, and annotated. Background proteins are depicted as gray dots and the respective bait protein annotated in red. **d** Scheme for the cloning of different fragments of TEBP-1, TEBP-2 and POT-1 for IP experiments and Y2H. TEBP-1 and TEBP-2 were divided into five fragments (f1–f5) of approx. 30 kDa, as well as two additional fragments covering the N-terminus including the predicted DNA-binding domains (f6) and the C-terminus (f7). POT-1 was divided into three fragments of around 15 kDa (f1–f3). **e** DNA pulldowns as in Fig. 1c with recombinantly expressed and N-terminally His-MBP-tagged fragments f1, f3, and f5 of TEBP-1 and TEBP-2, as well as the full length proteins with the same tags. The western blot was probed with α -His antibody and the signals detected by chemiluminescence. f1–f5: fragments of respective protein, full: full length respective protein, kDa: kilodalton, MBP: maltose-binding protein. $N = 2$ independent experiments with similar results. **f** Y2H assay like in (**a**) but with TEBP-1 and POT-1 full length proteins (fl), as well as N- and C-terminal fragments (f6 and f7 for TEBP-1, or f1 and f3 for POT-1, respectively) fused to the activation or DNA-binding domains of Gal4. Growth determined on the same medium as in **a**. **g** Y2H assay as in (**f**) but with TEBP-2 and POT-1 constructs. **h** Y2H assay as in (**f**) but with all fragments of TEBP-1 including the full length protein fused to the Gal4 DNA-binding domains, as well as all fragments of TEBP-2 including the full length protein fused to the Gal4 activation domain. f1–f7: fragments of respective protein, ctrl: control/empty plasmid, fl: full length protein. **i** Proposed working model for the interactions between telomere-binding proteins and telomere repeats in *C. elegans*. TEBP-1 and TEBP-2 fragments 3 (f3), containing a predicted DNA-binding domain, bind to ds telomere repeats and have opposing effects on telomere elongation. Both proteins interact with each other via their N-terminal fragments (f1). TEBP-1, TEBP-2 and POT-1 interact directly via the C-terminal fragment (f7) of TEBP-1/TEBP-2 and the N-terminal fragment (f1) of POT-1. As a result of this interaction, the ss telomere comes in closer contact to the ds telomere. Our current data does not support direct interactions between POT-1, POT-2, and MRT-1, but these factors may interact in the presence of telomeric DNA.

orthologs. The multiple sequence alignment showed the N-terminal region of *tebp* genes, the region with similarity to the homeodomains of human and yeast RAP1 (Supplementary Fig. 1d, e and Supplementary Data file 1), is more similar between orthologs than the C-terminal region (Supplementary Data File 2). However, phylogenetic analysis with only the N-terminal region did not produce major differences on tree topology (Supplementary Fig. 7). In order to derive evolutionary relationships between different *tebp* genes, we evaluated local synteny information. We found a high degree of regional synteny conservation between *C. elegans* *tebp-1* and one of the *tebp* copies in *C. inopinata*, *C. remanei*, *C. briggsae*, *C. nigoni*, *C. sinica*, *C. tropicalis*, and *C. japonica* (Table 1 and Supplementary Data file 2). Conversely, *tebp-2* did not show any signs of regional synteny across *Caenorhabditis* species (Supplementary Data file 2), suggesting that the gene duplication event creating *tebp-2* occurred after divergence from the *C. inopinata* lineage, less than 10.5 million years ago⁵³. Neither of the two *tebp* orthologs of *C. brenneri*, *C. latens*, and *C. angaria* are in synteny with *C. elegans* *tebp-1* (Supplementary Data file 2).

To determine whether TEBP proteins are generally telomere-binders in the *Elegans* supergroup, we performed DNA pulldowns, using nuclear extracts prepared from synchronized *C. briggsae* gravid adults. CBG11106, the only *C. briggsae* ortholog of *tebp-1* and *tebp-2*, was significantly enriched in the telomere pulldown (Fig. 7b), demonstrating that it can bind to the TTAGGC telomeric repeat. Of note, CBG22248, one of the two *C. briggsae* orthologs of MRT-1, was also enriched in the telomere pulldown, and CBG16601, the ortholog of POT-1, was just below our significance threshold, suggesting functional similarities to their *C. elegans* orthologs.

Discussion

Telomeres and their associated proteins are important to ensure proper cell division. In the popular model nematode *C. elegans*, only ss telomere-binding proteins were known thus far^{31,38}. Here, we describe a telomeric complex with the paralogs TEBP-1 and TEBP-2 as direct ds telomere-binding proteins. POT-1 seems to bridge the ds telomere-binding module of the complex, comprised of TEBP-1 and TEBP-2, with the ss telomere region. Strikingly, despite the high level of sequence similarity between TEBP-1 and TEBP-2, their mutant phenotypes are divergent.

Robust identification of telomere-associated proteins in *C. elegans*. Three lines of evidence demonstrate the validity and robustness of our screen. First, attesting for its technical reproducibility, the two qMS detection strategies employed shared an overlapping set of proteins enriched in telomeric sequence pulldowns (8 overlapping factors out of 12 and 8 hits). Second, within our overlapping set of enriched factors, we detected the previously identified ss telomere-binding proteins POT-1, POT-2, and MRT-1^{31,33,37,38}. Lastly, the *C. elegans* KU heterodimer homologs CKU-70 and CKU-80 were enriched in the screens. In other organisms, such as *Saccharomyces cerevisiae*, *Trypanosoma brucei*, *Drosophila melanogaster*, and *Homo sapiens*, KU proteins have been shown to associate with telomeres, regulating their length and protecting them from degradation and recombination^{54,55}. The *C. elegans* homologs were shown to interact with telomeres, but do not seem to have telomere regulatory functions⁴³. However, CKU-70 and CKU-80 were not enriched in the TEBP-1 and TEBP-2 interactome experiments, suggesting that their binding to telomeric DNA occurs independently of the TEBP-1/TEBP-2 complex (Fig. 5 and Supplementary Fig. 6). Alternatively, these factors may be part of the telomeric complex, with no direct interaction with TEBP-1 or TEBP-2.

We identified POT-3 in the background of our LFQ screen (Supplementary Data File 3), supporting the lack of telomeric phenotypes of *pot-3* mutants³¹. Furthermore, a number of factors previously reported to have telomere DNA-binding capability or to regulate telomere length, were not detected or lacked significant enrichment in our quantitative proteomics screen. MRT-2 is a homolog of *S. cerevisiae* checkpoint gene RAD17 and human RAD1, previously reported to regulate telomere length³⁰. Much like *tebp-2* and *mrt-1*, *mrt-2* mutants have shorter telomeres than wild-type and a Mrt phenotype. It is plausible that MRT-2 regulates telomere length beyond the context of direct telomeric binding. PLP-1⁵⁶, HMG-5⁵⁷, and CEH-37⁵⁸, were previously shown to bind to the *C. elegans* telomeric sequence in vitro. PLP-1 was enriched in the (AGGTCA)_n scrambled control in our qMS screen (Supplementary Data file 3), suggesting that PLP-1 is a general ds DNA binder, and not a specific telomere binder. Furthermore, HMG-5 was detected in the background, and CEH-37 was not detected altogether in our screen (Supplementary Data file 3). Further studies should clarify if and how these factors interact with the telomere complex described in this work.

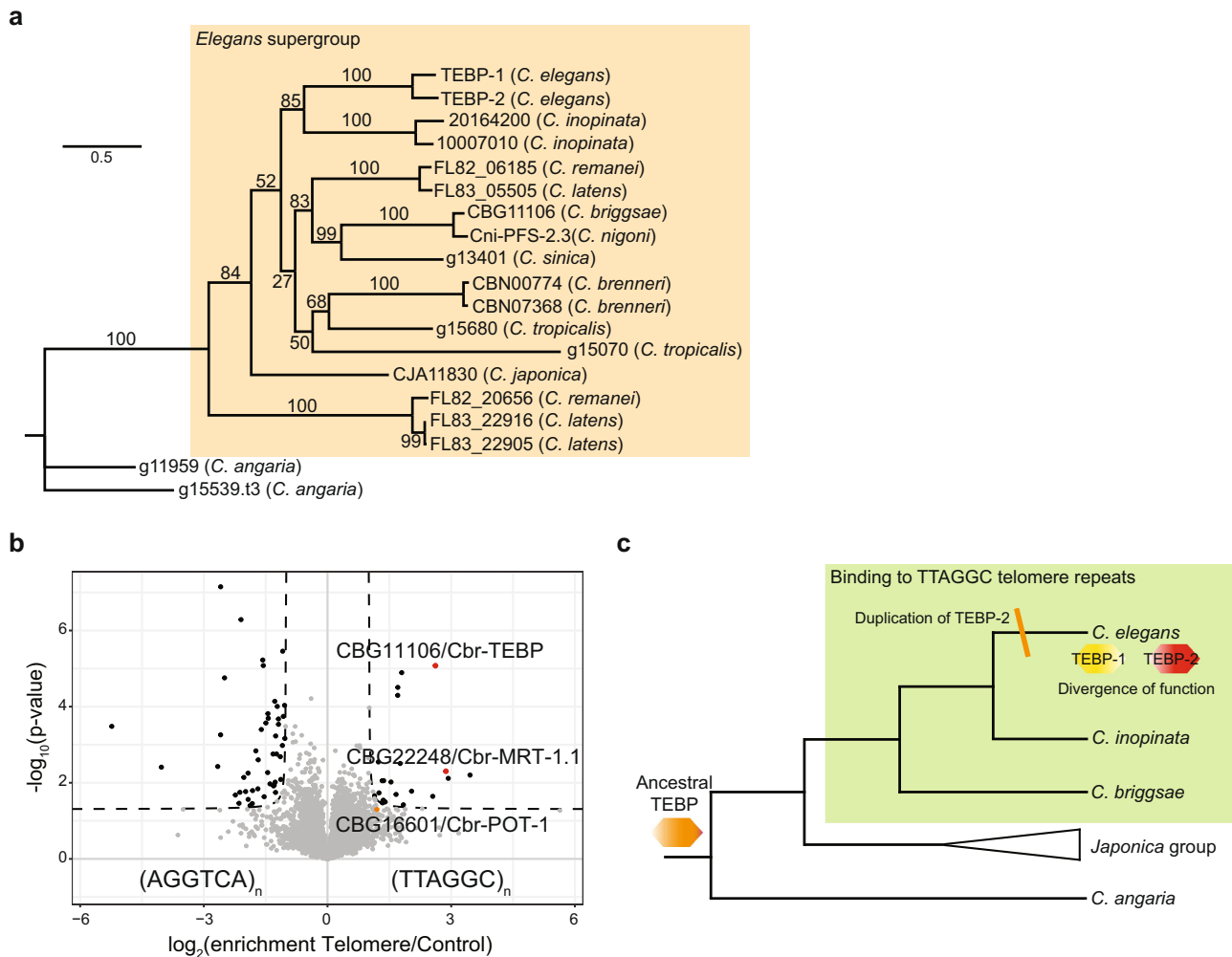


Fig. 7 Conservation of *tebp* genes in the *Caenorhabditis* genus. **a Phylogenetic tree constructed with IQ-TREE (v1.6.12), using a MAFFT (v7.452) multiple sequence alignment of the protein sequences of TEBP orthologs (see Supplementary Data file 2, sheet 2). Values on the nodes represent bootstrapping values for 10,000 replicates, set to 100. The TEBP orthologs outside the orange background represent the outgroup of the analysis. **b** Volcano plot of telomere DNA pulldown, as in Fig. 1a, of gravid adult nuclear extracts from *C. briggsae*. Here, pulldowns were performed in quadruplicates, per condition. Enriched proteins (enrichment threshold > 2-fold, p -value < 0.05) are labeled as black dots, whereas enriched proteins of interest are labeled with red or orange dots. Proteins below the threshold are depicted as gray dots. Homologs of telomere binders are named. **c** Depiction of the evolution of *tebp* genes in *Caenorhabditis*. We speculate that this family originated from an ancestor TEBP (orange hexagon), presumably required for fertility and capable of binding to telomeres. As we have confirmed telomere binding in *C. elegans* and *C. briggsae* (species in bold indicate confirmed binding of TEBP proteins to telomeric DNA), it is plausible that their common ancestor was able to bind to telomeres. The gene duplication that generated *tebp-2* occurred after the divergence of *C. elegans* and *C. inopinata* (marked as orange stripe), followed by division, or diversification, of functions of these two paralogs (TEBP-1: yellow hexagon, TEBP-2: red hexagon).**

The fast-evolving paralogs TEBP-1 and TEBP-2 are required for fertility. TEBP-1 and TEBP-2 share 65.4% of their amino acid sequence, which most likely reflects a common origin by gene duplication. Interestingly, the two paralogs TEBP-1 and TEBP-2 interact with each other, and with the same set of factors, i.e., POT-1, POT-2, and MRT-1 (Fig. 5 and Supplementary Fig. 6). This is striking, considering the divergent phenotypes of *tebp-1* and *tebp-2* mutants: *tebp-1* mutants have longer telomeres than wild-type, while *tebp-2* animals have shorter telomeres than wild-type and a Mortal Germline. Moreover, while the fertility of *tebp-1* and *tebp-2* animals is not compromised, *tebp-1*; *tebp-2* double mutants show highly penetrant synthetic sterility irrespective of the temperature the animals are grown at, indicating that TEBP-1 and TEBP-2 contribute to normal fertility (Fig. 4 and Supplementary Fig. 5). The observed synthetic sterility is likely justified by failure to enter and progress through normal mitosis and meiosis, as judged by the under-proliferation of germ cells.

The synthetic sterility of *tebp-1*; *tebp-2* animals is specific to these two paralogs, as other genetic crosses of shorter versus longer telomere mutants did not result in sterile double mutants. The synergistic role of TEBP-1 and TEBP-2 in fertility provide a puzzling contrast with their opposed telomere length mutant phenotypes. We speculate that the requirement of TEBP-1 and TEBP-2 to fertility may be independent of their functions at telomeres. Future studies on the influence of TEBP-1 and TEBP-2 on germline and embryonic gene expression may shed light on this aspect.

CBG11106, the single homolog of TEBP-1 and TEBP-2 in *C. briggsae*, interacts with telomeric DNA (Fig. 7b), suggesting that TEBP nematode homologs bind to telomeric DNA at least since the divergence of *C. elegans* and *C. briggsae*, from a common ancestor that presumably lived 80–100 million years ago⁵⁹. To verify this, the capability of additional TEBP orthologs to bind to telomeric DNA needs to be experimentally addressed. We

Table 1 Synteny analysis of *tebp* orthologs in other *Caenorhabditis* species.

<i>tebp</i> ortholog	Synteny with <i>tebp-1</i>	Synteny with <i>tebp-2</i>
10007010 (<i>C. inopinata</i>)	–	–
20164200 (<i>C. inopinata</i>)	+	–
FL82_06185 (<i>C. remanei</i>)	+	–
FL83_05505 (<i>C. latens</i>)	–	–
CBG11106 (<i>C. briggsae</i>)	+	–
Cni-PFS-2.3 (<i>C. nigoni</i>)	+	–
g13401 (<i>C. sinica</i>)	+	–
CBN00774 (<i>C. brenneri</i>)	–	–
CBN07368 (<i>C. brenneri</i>)	–	–
g15680 (<i>C. tropicalis</i>)	+	–
g15070 (<i>C. tropicalis</i>)	–	–
CJA11830 (<i>C. japonica</i>)	+	–
FL83_22916 (<i>C. latens</i>)	–	–
FL83_22905 (<i>C. latens</i>)	–	–
FL82_20656 (<i>C. remanei</i>)	–	–
g15539.t3 (<i>C. angaria</i>)	–	–
g11959 (<i>C. angaria</i>)	–	–

Overview of synteny of the *tebp* orthologs of other *Caenorhabditis* species with *tebp-1* or *tebp-2* of *C. elegans*. A “+” indicates regional synteny, while a “–” is lack of synteny.

speculate that *tebp-1* and *tebp-2* originated from an ancestor *Caenorhabditis* *tebp* gene required for fertility and with the ability to bind ds telomeric repeats (Fig. 7c). The *tebp-1* ancestor was duplicated after the divergence of *C. inopinata* and *C. elegans*, 10.5 million years ago⁵³, likely initiating a process of functional diversification of *tebp-1* and *tebp-2*.

Given their possible recent divergence, in evolutionary terms the 65.4 % protein sequence similarity observed between the protein sequences of TEBP-1 and TEBP-2 is actually fairly low. This likely reflects fast evolution of TEBP-1 and TEBP-2, in line with the known fast evolution as suggested for other telomere-binding proteins⁶⁰. While it is tempting to establish evolutionary relationships with vertebrate TRF1 and TRF2 proteins, TEBP-1/TEBP-2 and TRF1/TRF2 are not homologs. In addition, TRF1 and TRF2 are binding to telomeric DNA via C-terminal myb-domains⁶¹, while DNA binding in TEBP-1 and TEBP-2 occurs N-terminally. However, on the functional level, similarity between *C. elegans* TEBP-1/TEBP-2 and vertebrate TRF1/TRF2, potentially reflecting convergent evolution between two phylogenetically independent sets of telomere-binding paralogs is possible, but needs further investigation.

A telomere complex in actively dividing tissues in homeostasis.

Our size-exclusion chromatography, quantitative proteomics, and Y2H data support the existence of a telomere complex comprising TEBP-1, TEBP-2, POT-1, POT-2, and MRT-1 (Fig. 5 and Supplementary Fig. 6). According to our size-exclusion chromatography data, this complex elutes in a range between 600 kDa and 1.1 MDa. It should be noted that our model does not make any assumptions regarding complex stoichiometry. At the moment, we cannot exclude the existence of remaining DNA fragments in the complex, despite nuclease treatment, which could add to the total molecular weight. Thus, we propose a working model, whereby TEBP-1 and TEBP-2 bind to ds telomere repeats via their third predicted homeo-/myb-domains, and directly interact with the OB-fold of POT-1 with their C-terminal tails. Binding to POT-1 may, in turn, bring the ss telomeric repeats, and thus POT-2 and MRT-1, into closer contact (Fig. 6i). In the absence of POT-1, TEBP-1 and TEBP-2 are not able to interact with POT-2 and MRT-1 (Fig. 6b, c). We speculate that reciprocal regulation

by TEBP-2 and POT-1/TEBP-1 define normal telomere length. In this scenario, TEBP-2 might counteract telomere shortening by POT-1 and TEBP-1 (Fig. 6i). The precise interplay between these telomeric factors, namely the interactions between POT-1, POT-2, and MRT-1, and the mechanism of telomere elongation have to be further elucidated.

The mammalian shelterin complex counteracts recognition of telomeres as DNA double-strand breaks by inhibiting the DNA damage machinery. When shelterin factors are abrogated, catastrophic end-to-end chromosome fusions are observed^{62,63}. Previous studies did not identify end-to-end chromosome fusions in *pot-1* and *pot-2* mutants^{31,33,37}. It remains to be determined if *tebp-1* and/or *tebp-2* mutations lead to telomere fusions and whether the *C. elegans* telomeric complex is required to protect telomeres from DNA damage. It is possible that the synthetic sterility and high frequency of males observed in *tebp-1*; *tebp-2* double mutants, as well as the Mortal Germline phenotype of *tebp-2* and *mrt-1*, may be downstream of germline genome instability.

A germline-specific MAJIN/TERB1/TERB2 telomere-binding complex has been described in mouse testes^{64–66}. Knock-outs of these factors lead to meiotic arrest and male sterility^{64–66}, similar to the observed phenotype in *tebp-1*; *tebp-2* double mutants. This mammalian protein complex tethers telomeres to the nuclear envelope, a process essential for meiotic progression. A previous study has shown that POT-1 is required in *C. elegans* to tether telomeres to the nuclear envelope during embryogenesis⁶⁷. Given the interaction of TEBP-1 and TEBP-2 with POT-1 in vitro and in vivo, the telomeric complex may be dynamically involved in this process.

The distinct compartmentalization of post-mitotic soma versus actively dividing germline, together with a plethora of genetic tools, make *C. elegans* an enticing model organism for telomere biology in vivo, in homeostatic conditions. The identification of a telomeric complex in *C. elegans* allows further investigation of telomere regulation in this popular model organism.

Methods

C. elegans nuclear-enriched protein extract preparation. Nuclear extract preparation of gravid adult worms was done as described⁶⁸. The worms were synchronized by bleaching and harvested at the gravid adult stage by washing them off the plate with M9 buffer. After washing the worms in M9 buffer for 4 times, they were pelleted by centrifugation at 600 x g for 4 min, M9 buffer was removed and extraction buffer (40 mM NaCl, 20 mM MOPS pH 7.5, 90 mM KCl, 2 mM EDTA, 0.5 mM EGTA, 10% Glycerol, 2 mM DTT, and 1x complete protease inhibitors Roche) was added. Worms resuspended in extraction buffer were frozen in liquid nitrogen. The resulting pellets were ground to a fine powder in a pre-cooled mortar and transferred to a pre-cooled glass douncer. When thawed, the samples were sheared with 30 strokes, piston B. The worm suspension was pipetted to pre-cooled 1.5 ml reaction tubes (1 ml per tube) and cell debris, as well as unshredded worms were pelleted by centrifugation at 200 x g for 5 min at 4 °C for two times. To separate the cytoplasmic and nuclear fractions, the supernatant was spun at 2000 x g for 5 min at 4 °C. The resulting pellet containing the nuclear fraction was washed twice by resuspension in extraction buffer and subsequent centrifugation at 2000 x g for 5 min at 4 °C. After the washing steps, the nuclear pellet was resuspended in 200 µl buffer C + (420 mM NaCl, 20 mM Hepes/KOH pH 7.9, 2 mM MgCl₂, 0.2 mM EDTA, 20% Glycerol, 0.1% Igepal CA 630, 0.5 mM DTT, 1x complete protease inhibitors). Nuclear extract of gravid adult worms of *C. briggsae* was prepared as described above.

Oligonucleotides. All oligonucleotides used throughout this manuscript (cloning, sequencing, DNA pulldowns, fluorescence polarization etc.) are listed in Supplementary Data file 4 with their name and sequence.

DNA pulldowns

Preparation of biotinylated DNA for pulldown experiments. Biotinylated telomeric and control DNA for the DNA pulldown for detection of telomeric interactors was prepared as previously published^{16,39,40}. In short, 25 µl of 10-mer repeat oligonucleotides of either telomeric or control sequence were mixed 1:1 with 25 µl of their respective reverse complement oligonucleotide and 10 µl annealing buffer (200 mM Tris-HCl, pH 8.0, 100 mM MgCl₂, 1 M KCl). The mixture was brought to

100 μ l final volume with H₂O, heated at 80 °C for 5 min, and left to cool. Once at room temperature (RT), the samples were supplemented with 55 μ l H₂O, 20 μ l 10x T4 DNA ligase buffer (Thermo Scientific), 10 μ l PEG 6000, 10 μ l 100 mM ATP, 2 μ l 1 M DTT and 5 μ l T4 Polynucleotide Kinase (NEB, 10 U/ μ l, #M0201) and left at 37 °C for 2 h to concatenate. Finally, 4 μ l of T4 DNA Ligase (Thermo Scientific, 5 WU/ μ l, #EL0011) were added and the samples incubated at RT overnight for ligation and polymerization. The ligation process was monitored by running 1 μ l of the reaction on a 1% agarose gel. The samples were cleaned by phenol-chloroform extraction. For this, 1 vol. of H₂O and 200 μ l of Phenol/Chloroform/Isoamyl Alcohol (25:24:1; pH 8; Invitrogen, # 15593049) was added to the mixture, vortexed and centrifuged at 16,000 xg for 2 min. After centrifugation the aqueous phase was transferred to a fresh tube and the DNA precipitated by addition of 1 ml 100% Ethanol and incubation at -20 °C for 30 min. Afterwards the suspension was centrifuged at 16,000 xg for 45 min at 4 °C. The resulting DNA pellet was resuspended in 74 μ l H₂O and 10 μ l 10x Klenow-fragment reaction buffer (Thermo Scientific), 10 μ l 0.4 mM Biotin-7-dATP (Jena Bioscience, #NU-835-BIO) and 6 μ l Klenow-Fragment exo- polymerase (Thermo Scientific, 5 U/ μ l, # EP0422) added. Biotinylation was carried out by incubation at 37 °C over night. The reaction was cleaned up by size-exclusion chromatography using MicroSpin Sephadex G-50 columns (GE Healthcare, #GE27-5330-01).

Pull-down experiments. Biotinylated DNA and Dynabeads™ MyOne™ Streptavidin C1 (Thermo Scientific, #65001) were mixed with PBB buffer (50 mM Tris/HCl pH 7.5, 150 mM NaCl, 0.5% NP 40, 5 mM MgCl₂, 1 mM DTT) and incubated at room temperature for 15 min on a rotating wheel to immobilize the DNA on the beads. After three washes with PBB buffer, the DNA coupled beads were resuspended in PBB buffer and Salmon sperm (10 mg/ml, Ambion, #AM9680) was added 1:1000 as competitor for unspecific DNA binding. The pull-downs were performed with different amounts of protein extract (see below) and incubated at 4 °C on a rotating wheel for 90 min. Following incubation the beads were washed three times with PBB buffer and resuspended in 1x Loading buffer (4x NuPAGE LDS sample buffer, Thermo Scientific, #NP0008) supplemented with 100 mM DTT. For elution, the samples were boiled at 70 °C for 10 min and afterwards loaded on a gel and processed as indicated above for MS, or below for western blot. In pull-down-MS experiments, the pull-downs were prepared in either technical quadruplicates (LFQ), or technical duplicates (DML) per condition, whereas for western blot all conditions were prepared with one replicate and an input. In all, 200–400 μ g of nuclear worm extract and of *Escherichia coli* extract were used for the mass spectrometry screen and pull-downs of Fig. 1c, respectively. In all, 0.4–0.7 mg of total protein extract were used for the pull-downs shown in Fig. 1d–f. Four-hundred micrograms of *E. coli* extract was used in DNA-binding domain pull-downs in Fig. 6e.

Mass spectrometry: sample preparation, data acquisition, and analysis

In-gel digest. In-gel digestion was performed as previously described^{16,69} with the exception of the DML samples (see below). Samples were run on a 10% Bis-Tris gel (NuPAGE; Thermo Scientific, #NP0301) for 10 min (IP samples) or on a 4–12% Bis-Tris gel (NuPAGE, Thermo Scientific, #NP0321) for 20 min (LFQ-measured telomeric DNA pull-downs) at 180 V in 1x MOPS buffer (NuPAGE, Thermo Scientific, #NP0001). Individual lanes were excised and cut to approximately 1 mm × 1 mm pieces with a clean scalpel, and transferred to a 1.5 ml tube. For the LFQ telomeric DNA pull-downs, the lanes were split into four fractions. The gel pieces were destained in destaining buffer (50% 50 mM NH₄HCO₃ (ABC), 50% ethanol p.a.) at 37 °C under rigorous agitation. Next, gel pieces were dehydrated by incubation in 100% acetonitrile for 10 min at 25 °C shaking and ultimately dehydrated using a Concentrator Plus (Eppendorf, #5305000304, settings V-AQ). The gel pieces were incubated in reduction buffer (50 mM ABC, 10 mM DTT) at 56 °C for 60 min and subsequently incubated in alkylation buffer (50 mM ABC, 50 mM iodoacetamide) for 45 min at room temperature in the dark. Gel pieces were washed in digestion buffer (50 mM ABC) for 20 min at 25 °C. Next, gel pieces were dehydrated again by incubation in 100% acetonitrile and drying in the concentrator. The dried gel pieces were rehydrated in trypsin solution (50 mM ABC, 1 μ g trypsin per sample, Sigma-Aldrich, #T6567) and incubated overnight at 37 °C. The supernatant was recovered and combined with additional fractions from treatment with extraction buffer (30% acetonitrile) twice and an additional step with pure acetonitrile for 15 min at 25 °C, shaking at 1400 rpm. The sample solution containing the tryptic peptides was reduced to 10% of the original volume in a Concentrator Plus, to remove the acetonitrile and purified using the stage tip protocol.

Dimethyl labeling. Dimethyl labeling (DML) was done as previously described⁷⁰. For DML, in-gel digest was performed as indicated in the last section, with the exception of exchanging ABC buffer for 50 mM TEAB (Fluka, #17902) after alkylation. The volume of the extracted peptides was reduced in a Concentrator Plus. For labeling, either 4% formaldehyde solution (Sigma-Aldrich, #F8775) for light labeling or 4% formaldehyde-D2 (Sigma-Aldrich, #596388) solution for medium labeling, as well as 0.6 M NaBH₃CN (Sigma-Aldrich, #156159) were added to the samples and mixed briefly. The mixture was incubated for 1 h at 20 °C, shaking at 1000 rpm and afterwards quenched by addition of a 1% ammonia solution (Sigma-Aldrich, #30501) and acidified with 10% formic acid solution (Merck, #1.00264.1000). After the labeling reaction, the respective light and

medium samples were mixed 1:1 (light telomere: medium control; medium telomere: light control) and purified by stage tip purification.

Stage tip purification. Stage tip purification was performed as previously described⁷¹. Desalting tips were prepared by using two layers of Empore C18 material (3 M, #15334911) stacked in a 200 μ l pipet tip. The tips were activated with pure methanol. After two consecutive washes with Buffer B (80% acetonitrile, 0.1% formic acid) and Buffer A (0.1% formic acid) for 5 min the tryptic peptide samples were applied and washed once more with Buffer A. Upon usage, peptides were eluted with Buffer B. The samples were centrifuged in a Concentrator Plus for 10 min to evaporate the acetonitrile and adjusted to 14 μ l with Buffer A.

MS measurement and data analysis. For MS measurement 5 μ l of sample were injected. The desalted and eluted peptides were loaded on an in-house packed C18 column (New Objective, 25 cm long, 75 μ m inner diameter) for reverse-phase chromatography. The EASY-nLC 1000 system (Thermo Scientific) was mounted to a Q Exactive Plus mass spectrometer (Thermo Scientific) and peptides were eluted from the column in an optimized 2 h (pulldown) gradient from 2 to 40% of 80% MS grade acetonitrile/0.1% formic acid solution at a flow rate of 225 nL/min. The mass spectrometer was used in a data-dependent acquisition mode with one MS full scan and up to ten MS/MS scans using HCD fragmentation. All raw files were processed with MaxQuant (version 1.5.2.8) and searched against the *C. elegans* Wormbase protein database (Version WS269), as well as the Ensembl Bacteria *E. coli* REL606 database (version from September 2018) for proteins from the feeding strain OP50. Carbamidomethylation (Cys) was set as fixed modification, while oxidation (Met) and protein N-acetylation were considered as variable modifications. For enzyme specificity, trypsin was selected with a maximum of two miscleavages. LFQ quantification (without fast LFQ) using at least 2 LFQ ratio counts and the match between run option were activated in the MaxQuant software. Fractions and conditions were indicated according to each experiment. Data analysis was performed in R using existing libraries (ggplot2-v 3.2.1, ggrepel-v 0.8.1, stats-v 3.5.2) and in-house scripts. Protein groups reported by MaxQuant were filtered removing known contaminants, protein groups only identified by site and those marked as reverse hits. Missing values were imputed at the lower end of LFQ values using random values from a beta distribution fitted at 0.2–2.5%. For statistical analysis, *p*-values were calculated using Welch's *t*-test. Enrichment values in the volcano plots represent the mean difference of log₂ transformed and imputed LFQ intensities between the telomere and the control enriched proteins. Peptide labels created by the dimethyl-labeling reaction were selected in the MaxQuant software as “N-terminal Dimethyl 0” and “Dimethyl 0” for the light samples, as well as “N-terminal Dimethyl 4” and “Dimethyl 4” for the heavy labeled samples. The re-quant option was activated. An incorporation check was run additionally to confirm incorporation of the dimethyl labels of at least 95% in each sample. Protein groups resulting from MaxQuant analysis were filtered identically to LFQ. The normalized ratios for each protein were log₂ transformed and plotted in the scatterplot. Filtering and analysis were done in R using existing libraries and an in-house script.

In vitro single- or double-strand binding of proteins from *C. elegans* extract.

For this assay, biotinylated oligonucleotides (Metabion) were used, containing a five times repeat of telomeric G-rich, C-rich, or control sequences. To allow for proper annealing, all oligonucleotides contained unique sequences flanking both sides of the repeats. Double-stranded oligonucleotides were prepared by mixing the biotinylated forward oligonucleotide 1:1 with the respective non-biotinylated reverse complement oligonucleotide and addition of annealing buffer (200 mM Tris-HCl, pH 8.0, 100 mM MgCl₂, 1 M KCl). The mix was heated at 80 °C for 5 min and cooled to room temperature. The single-stranded oligonucleotides were treated similarly, only replacing the reverse complement oligonucleotide with H₂O. The pulldown itself was performed as described above with 0.5 mg (TEBP-2::GFP) or 0.4 mg (TEBP-1::3xFLAG) *C. elegans* embryo total protein extract of the respective strains. After elution, the samples were run on a 4–12% Bis-Tris gel (NuPAGE, Thermo Scientific, #NP0321) at 150 V for 120 min and transferred to a membrane. Western blot detection of the tagged proteins was carried out as described below.

Expression and purification of recombinant protein from *E. coli*. Auto-induction⁷² was used for expression of His₆-MBP-POT-2. An overnight culture of the expression strain BL21(DE3) was cultured at 37 °C in YG medium (2% Yeast extract, 0.5% NaCl, 3.5% Glycerol) supplemented with the respective antibiotic. A growing culture in YG medium was prepared by inoculating it with 1:50 volume of the overnight culture. At an OD₆₀₀ of 0.7, a culture of auto-induction medium (2% Peptone, 3% Yeast extract, 25 mM Na₂HPO₄/KH₂PO₄, 0.05% Glucose, 2.2% Lactose, 0.5% Glycerin, 50 mM NH₄Cl, 5 mM Na₂SO₄, 2 mM MgSO₄, 1x Trace Metal Solution) was inoculated with the growing culture to a density of OD₆₀₀ 0.004. 1000x Trace Metal Solution used for the auto-induction medium, has the following constitution: of 50 mM FeCl₃/HCl, 20 mM CaCl₂, 10 mM Mn(II)Cl₂, 10 mM ZnCl₂, 2 mM CoCl₂, 2 mM Cu(II)Cl₂, 2 mM NiCl₂, 2 mM NaMoO₄, 2 mM Na₂SeO₃. The auto-induction culture was incubated at 25 °C for 24 h and then harvested by centrifugation at 4000 xg.

TEBP-1-His₅ and TEBP-2-His₅ were expressed in Rosetta 2 (DE3) pLysS competent cells (Novagen, #71401). An overnight culture was grown in LB containing the respective antibiotic. A growing culture was inoculated and after reaching mid-log growth at 37 °C, the cultures were induced with 1 mM IPTG. Cells were grown at 18 °C and harvested after 24 h. IPTG-induced or auto-induction cultures were pelleted in 50 ml reaction tubes by centrifugation at 4000 x g after growth and lysed according to the protocol for the respective downstream use.

POT-2 expression pellets were resuspended in Tris buffer (50 mM Tris/HCl pH 7.5, 100 mM NaCl, 10 mM MgCl₂, 1x EDTA-free protease inhibitor (Roche, #4693132001)) and divided into 2 ml flat lid micro tubes containing 0.1 mm zirconia beads (Carl Roth, #N033.1). Lysis of the cells was achieved with a FastPrep -24™ Classic (MP Biomedicals, #116004500) using the setting 6 m/s for 30 s for two times. In between the disruption cycles the samples were centrifuged at 21,000 x g for 2 min to pellet debris, followed by an incubation on ice for 5 min before the second cycle. After lysis the suspension was centrifuged at 21,000 x g for 10 min at 4 °C.

TEBP-1 and TEBP-2 expression pellets were lysed via sonication with a Branson Sonifier 450 (duty cycle: 50%, output control: 3, 3.5 min with 5 mm tip) in lysis buffer (25 mM Tris-HCl pH 7.5, 300 mM NaCl, 20 mM imidazole) with 1 mM DTT, and protease inhibitor cocktail tablets (Roche, #4693132001). Lysates were centrifuged at 4613 x g for 10 min at 4 °C. For both preparation methods the supernatant was afterwards transferred to fresh reaction tubes.

His-MBP tagged TEBP-1 and TEBP-2 fragments were expressed in E.coli ArcticExpress DE3 cells (Agilent, #230192). Cultures were grown overnight in 5 ml LB supplemented with the respective antibiotic for the expression vector. Next day the expression culture was inoculated from the overnight culture and grown to mid-log phase at 30 °C, and then induced with 1 mM IPTG. Cultures were incubated at 12 °C and harvested after 24 h. The pellet was resuspended in binding buffer (20 mM Tris-HCl pH 7.5, 500 mM NaCl, 50 mM imidazole) with 1 mM DTT, complete protease inhibitor cocktail tablets (Roche, #4693132001), and 100 µg DNase I (NEB, M0303S). Cells were lysed using a Branson Sonifier (duty cycle: 50%, output control: 4, 6 min (3 min sonication, 3 min ice, 3 min sonication) with 9 mm tip). Lysates were cleared at 4613 x g for 10 min at 4 °C, and used for subsequent assays.

Protein expression, purification, and fluorescence polarization assay. E.coli ArcticExpress DE3 cells (Agilent, #230192) were grown overnight in 5 ml LB supplemented with the respective antibiotic for the expression vector. Next day the expression culture was inoculated from the overnight culture and grown to mid-log phase at 30 °C, and then induced with 1 mM IPTG. Cultures were incubated at 12 °C and harvested after 24 h. The pellet was resuspended in binding buffer (20 mM Tris-HCl pH 7.5, 500 mM NaCl, 50 mM imidazole) with 1 mM DTT, complete protease inhibitor cocktail tablets (Roche, #4693132001), and 100 µg DNase I (NEB, M0303S). Cells were lysed using a Branson Sonifier (duty cycle: 50%, output control: 4, 6 min (3 min sonication, 3 min ice, 3 min sonication) with 9 mm tip). Lysates were ultracentrifuged (Beckman Optima XE-100) at 75,000 x g for 30 min at 4 °C. After loading the lysate, the HisTrap HP column (GE Healthcare, #GE17-5247-01) was washed with binding buffer, and proteins were eluted in binding buffer containing 500 mM imidazole in 250 µl fractions. Proteins were dialyzed with the PD-10 Desalting Column (GE Healthcare, #GE17-0851-01) in a buffer consisting of 20 mM Tris-HCl pH = 7.5, 1 mM MgCl₂, 150 mM NaCl, 10% (v/v) glycerol, and 1 mM DTT, and were concentrated. These fractions were then utilized for the fluorescence polarization assays.

The purified protein stocks were used from a maximum concentration of 4 µM, to a minimum concentration of 2 nM in twofold serial dilutions in ice-cold buffer containing 20 mM HEPES pH 7.0, 100 mM NaCl, and 5% (v/v) glycerol. FITC-labeled oligonucleotides (Metabion) carrying 2.5x, 2.0x, and 1.5x repeats of either telomeric (G- or C-rich), or control sequence were used for this assay. Double-stranded oligonucleotides were prepared by mixing 1:1 with the respective reverse complement oligonucleotide. For annealing, oligonucleotides were heated to 95 °C and then cooled at 0.1 °C/s until 4 °C. Diluted proteins were incubated with a final concentration of 20 nM FITC-labeled probe for 10 min at room temperature. Samples were measured with a Tecan Spark 20 M (Tecan). Experiments were conducted using three replicates for each condition. Analysis was performed with Graph Pad Prism 9.0 and specific binding was measured with Hill slope.

C. elegans complete protein extract preparation. Animals were washed off the plates with M9 buffer, synchronized by bleaching and grown to the desired stage, at which point worms were collected with M9 buffer. Worms were washed 3–4 times in M9, washed one last time with H₂O and frozen in 100–200 µl aliquots. Upon extract preparation, the aliquots were thawed, mixed 1:1 with 2x Lysis Buffer (50 mM Tris/HCl pH 7.5, 300 mM NaCl, 3 mM MgCl₂, 2 mM DTT, 0.2 % Triton X-100, Protease inhibitor tablets), and sonicated in a Bioruptor 300 (Diagenode) for 10 cycles with 30 s on/off, on high level. After sonication, the samples were centrifuged at 21,000 x g for 10 min to pellet cell debris. The supernatant was transferred to a fresh tube. With the exception of embryos (see below), extract of all developmental stages of *C. elegans* was prepared as described above. Samples of each developmental stage (for Fig. 2a) were collected in the following time points after plating of synchronized L1s: L1s were collected ~7 h after plating to recover

from starvation; L2s, ~12 h; L3s, ~28 h; L4, ~49 g; and YAs were collected ~ 56 h after plating.

For mixed-stage embryo extract preparations, synchronized gravid adults were harvested by washing them off the plate with M9 buffer. The worm suspension was washed with M9 until the supernatant was clear. Then, animals were bleached until all gravid adults were dissolved and only mixed-staged embryos remained. The embryos were subsequently washed in M9 buffer for three times then transferred to a new tube and washed one more time. In the last wash step the embryos were resuspended in 1x lysis buffer (25 mM Tris/HCl pH 7.5, 150 mM NaCl, 1.5 mM MgCl₂, 1 mM DTT, 0.1 % Triton X-100, protease inhibitors) and frozen in liquid nitrogen. After freezing, the pellets were ground to a fine powder in a pre-cooled mortar, then transferred to a cold glass douncer and sheared for 40 strokes with piston B. The suspension was pipetted to 1.5 ml tubes and spun down at 21,000 x g for 15 min at 4 °C. Finally, the supernatant was transferred to a fresh tube.

Immunoprecipitation (IP)

GFP IP. IPs with GFP-tagged proteins were performed with GFP-binding magnetic agarose beads (GFPtrap MA, Chromotek, #gtma-20). Per IP sample, 10 µl of bead slurry was used and washed two times with 500 µl Wash Buffer (10 mM Tris/HCl pH 7.5, 150 mM NaCl, 0.5 mM EDTA, 1:1000 Pepstatin A/Leupeptin, 1:100 PMSF). Afterwards, the beads were resuspended in 450 µl Wash Buffer and up to 1 mg of complete extract of the respective *C. elegans* strain (of mixed-stage embryos or young adults) was added to a final volume between 500 and 750 µl. The IP samples were incubated at 4 °C rotating for 2 h. Following three washing steps with 500 µl Wash Buffer the beads were resuspended in 1x LDS (4x NuPAGE LDS sample buffer, Thermo Scientific, #NP0008) supplemented with 100 mM DTT and boiled at 70 °C for 10 min. When used for mass spectrometry, the samples were prepared in quadruplicates per strain/condition. In the IP-MS related to Supplementary Fig. 5e, f, the Wash Buffer was supplemented with 2 mM MgCl₂ and 0.05% of recombinant endonuclease from *Serratia marcescens*, or Sm nuclease⁷³, produced by the IMB's Protein-Production Core Facility.

FLAG IP. IPs with FLAG-tagged protein were performed with Protein G magnetic beads (Invitrogen™ Dynabeads™ Protein G; #10004D) and α-FLAG antibody (Monoclonal ANTI-FLAG® M2 antibody produced in mouse, Sigma Aldrich, #F3165). Per IP, 30 µl of beads were used and washed three times with 1 ml Wash Buffer (25 mM Tris/HCl pH 7.5, 300 mM NaCl, 1.5 mM MgCl₂, 1 mM DTT, 1 complete Mini protease inhibitor tablet per 50 ml). The beads were resuspended in 450 µl Wash Buffer and up to 1 mg of complete protein extract from the respective *C. elegans* strains was added. Finally, 2 µg of FLAG antibody were added and the samples were incubated for 3 h, rotating at 4 °C. After the incubation, the samples were washed three to five times with 1 ml Wash Buffer (see washing steps before), the beads were resuspended in 1x LDS/DTT, and the samples were boiled at 95 °C for 10 min. For mass spectrometry, IPs were prepared in quadruplicates per strain/condition. When doing the IP with Sm nuclease, the wash buffer was supplemented with 0.05% Sm nuclease (as indicated above).

Western blot. Protein samples were boiled at 70 °C for 10 min and loaded on a 4–12% Bis-Tris gel (NuPAGE, Thermo Scientific, #NP0321), running at 150–180 V for 60–120 min in 1x MOPS. After the run, the gel was shortly washed in VE H₂O and equilibrated in transfer buffer (25 mM Tris, 192 mM Glycine, 20% Methanol). A nitrocellulose membrane (Amersham Protran, VWR, #10660002) was equilibrated in transfer buffer as well. Membrane and gel were stacked with pre-wet Whatman paper (GE Healthcare-Whatman, #WHA10426892) and immersed in a blotting tank (Bio-Rad) filled with ice-cold transfer buffer and additionally cooled with a cooling element. The proteins were blotted at 300 mA for 60–120 min depending on the size. If blotted for 90–120 min for larger proteins, the transfer was carried out with a blotting tank on ice to keep the temperature. After blotting, the membranes were further prepared according to the respective antibody protocol.

Anti-His antibody. Membranes were blocked in Blocking Solution (PentaHis Kit, Qiagen, #34460) for 1 h at room temperature. After three 5 min washes in TBS-T (1x TBS, 0.1% Tween-20, 0.5% Triton X-100) the membranes were incubated with the Anti-His-HRP conjugated antibody in a dilution of 1:1000 in Blocking Solution for 1 h at room temperature. The membranes were then washed again three times in TBS-T and incubated with ECL Western Blot reagent (Thermo Scientific™ SuperSignal™ West Pico PLUS Chemiluminescent Substrate, #15626144; mixed 1:1) for detection. Western blot ECL detection was performed with the ChemiDoc XRS+ system (BioRad, Software: Image Lab 5.2.1).

Anti-GFP, Anti-FLAG, and Anti-Actin antibodies. Western blot analysis was performed using the following primary antibodies: an anti-GFP antibody (Roche, Anti-GFP from mouse IgG1k (clones 7.1 and 13.1), #11814460001; 1:1000 in Skim Milk solution), an anti-FLAG antibody (Sigma-Aldrich, mouse Monoclonal ANTI-FLAG® M2 antibody, # F3165; 1:5000 in Skim Milk solution), and an anti-Actin antibody (Sigma-Aldrich, rabbit anti-actin, #A2066; 1:500 in Skim Milk solution). After blotting, membranes were blocked in Skim Milk solution (1x PBS, 0.1% Tween-20, 5% (w/v) Skim Milk Powder) for 1 h at room temperature. The

incubation with the primary antibody was carried out at 4 °C, rotating overnight. Membranes were washed in PBS-T (1x PBS, 0.1% Tween-20) three times for 10 min, they were incubated with an HRP-linked secondary antibody (for anti-flag and anti-GFP with Cell Signaling Technology, anti-mouse IgG, #7076; 1:10,000 dilution in Skim Milk Solution; for anti-actin the secondary used was GE Healthcare, anti-Rabbit IgG, #NA934; 1:3000 in Skim Milk solution) for 1 h rotating at room temperature. Following three washes in PBS-T the membranes were incubated with ECL solution (Thermo Scientific™ SuperSignal™ West Pico PLUS Chemiluminescent Substrate, #15626144; mixed 1:1) for detection. Western blot ECL detection was performed with the ChemiDoc XRS+ system (BioRad, Software: Image Lab 5.2.1). Incubation with Anti-Actin antibody was typically performed after detection of GFP/FLAG and subsequent washes.

Antibody protocol for co-IPs (LI-COR antibodies). For co-IP experiments, we first probed the IP bait with HRP-linked secondary antibodies, as described above. Then, we probed for the co-IP using LI-COR secondary antibodies. After incubation with primary antibody, as described above, membranes were washed and incubated with secondary antibodies compatible with the LI-COR System (FLAG/GFP: Licor IRDye® 680RD Donkey anti-Mouse IgG (H+L), #926-68072; Actin: Licor IRDye® 800CW Donkey anti-Rabbit IgG (H+L), #926-32213; both 1:15,000 in Skim Milk solution) for 1 h at room temperature. After three additional washes with PBS-T, the membranes were imaged using an Odyssey CLx scanner and processed using Image Studio software (LI-COR, Version 3.1).

C. elegans culture and strains. *C. elegans* was cultured under standard conditions on Nematode Growth Medium (NGM) plates seeded with *E. coli* OP50 bacteria⁷⁴. For proteomics experiments, animals were grown on OP50 high-density plates (adapted from ref.⁷⁵). In specific, the yolks of commercially available chicken eggs were isolated, added to LB medium (50 ml per egg yolk) and thoroughly mixed. Subsequently, the mix was incubated at 65 °C for 2–3 h. Pre-grown OP50 liquid culture is added to the mix (10 ml per egg), after the yolk-LB mixture cooled down. This preparation was poured into 9 cm plates (10 ml per plate) and plates are decanted the next day. Plates remained for 2–3 days at room temperature, for further bacterial growth and drying.

Animals were grown at 20 °C, except when noted. The standard wild-type strain used in this study was N2 Bristol. Strains used and created in this study are listed in Supplementary Table 1.

Fertility assays. For brood size counts of the homozygous single mutants, L3 worms were isolated, per strain and were grown either at 20 or 25 °C. After reaching adulthood, worms were transferred to a new plate every day, until no eggs were laid in 2 consecutive days. Viable progeny was counted approximately 24 h after removing the parent. For the experiment shown in Fig. 4d, e, a cross between *tebp-1(xf133)* males and *tebp-2(xf131)* hermaphrodites was performed, the genotypes of the F1 and F2 were confirmed by PCR genotyping. L2/L3 progeny of F2 *tebp-1(xf133)/+; tebp-2(xf131)* mothers were isolated and grown at 20 °C, or 25 °C. During adulthood, the viable brood size was counted as mentioned above. The assayed F3s were genotyped 2 days after egg laying stopped. For all brood size experiments, worms that died before egg laying terminated, e.g., by dehydration on the side of plate, were excluded from the analysis.

Mortal germline assay. All strains used in the Mortal Germline assay were outcrossed with wild-type N2 two times before the experiment. Six L3 larvae of the chosen strains were picked per plate ($n = 15$ plates per strain) and grown at 25 °C. Six L3 larvae were transferred to a fresh plate every 5 days (equivalent to two generations). This procedure was followed until plates were scored as sterile, when the six worms transferred failed to produce six offspring to further isolate, on 2 consecutive transfer days.

***pgl-1::mTagRfp-T; tebp-1 x pgl-1::mTagRfp-T; tebp-2* cross and definition of categories of germline defects.** We crossed *pgl-1::mTagRfp-T; tebp-1* males with *pgl-1::mTagRfp-T; tebp-2* hermaphrodites. F1 cross progeny was confirmed by genotyping. 300 F2 progeny were singled and left to self-propagate. After genotyping F2 worms, we isolated 60 F3 worms from three different *tebp-1(xf133);tebp-2(xf131)/+*, 60 F3 worms from three different *tebp-1(xf133)/+;tebp-2(xf131)* mothers, as well as 10 F3 worms from two different single mutant mothers as controls. Additionally, all synthetic sterility escaper progeny from *tebp-1; tebp-2* double-homozygous worms were singled to check their fertility. Germline health, as well as growth and other phenotypes for all singled worms were determined at day 2 of adulthood. Germlines were categorized by microscopy with a Leica M80 stereomicroscope with a fluorescence lamp (Leica EL 6000), according to the morphology of the germline, as assessed by PGL-1::mTagRFP-T expression: category 1, near wild-type morphology; category 2, one gonad arm is atrophied; category 3, both gonad arms are atrophied. After germline categorization, worms were genotyped. We repeated this procedure until the F5, always using the progeny of *tebp-1(xf133);tebp-2(xf131)/+* or *tebp-1(xf133)/+;tebp-2(xf131)* mothers, as well as sibling controls. The barplots depicting the final distribution of germline categories across all scored generations was created using R and publicly available packages (ggplot2-v 3.2.1, reshape-v 0.8.8, viridis-v 0.5.1, scales-v 1.0.0).

Scoring crosses of *tebp-1 x tebp-2* mutant animals. Owing to the onset of synthetic sterility in F2 *tebp-1; tebp-2* double mutant animals, > 100 of F2 progeny was singled from the F1 heterozygous parent. F2 worms were genotyped at the adult stage after 3–4 days of egg laying and genotypes were determined and correlated with fertility. Progeny descending from *tebp-1; tebp-2* double mutant synthetic sterility escaper F2s were singled and allowed to grow and lay eggs for 3–4 days. Subsequently, these double mutant F3s were genotyped and their fertility was determined. Boxplots depicting the results were created using R and publicly available packages (ggplot2-v 3.2.1, reshape-v 0.8.8, viridis-v 0.5.1, scales-v 1.0.0).

Creation of mutants using CRISPR-Cas9 technology. Mutants were created as described⁷⁶, with the following specifications. To create *tebp-2(xf131)*, N2 animals were injected with a mix of three constructs: 25 ng/μl of co-injection marker pCFJ104 (*Pmyo-3:mCherry:unc-54 3'UTR*, a gift from Erik Jorgensen, Addgene plasmid #19328; <http://n2t.net/addgene:19328>; RRID:Addgene_19328); 100 ng/μl of a construct expressing Cas9 and a sgRNA targeting the sequence ACAT-GAGTCTGTGTTACGG (derived from pDD162, which was a gift from Bob Goldstein, Addgene plasmid # 47549; <http://n2t.net/addgene:47549>; RRID: Addgene_47549); and 75 ng/μl of a construct expressing a sgRNA targeting ACGGCTCATAAGAGACTTGG (derived from p46169, which was a gift from John Calarco, Addgene plasmid # 46169; <http://n2t.net/addgene:46169>; RRID: Addgene_46169).

To produce *tebp-1(xf133)* and *tebp-1(xf134)*, the following mix was injected into N2 animals: 25 ng/μl of pCFJ104; 150 ng/μl of a construct expressing Cas9 and a sgRNA targeting the sequence GCATGTCGAGATTCTACTGG (derived from pDD162); and 80 ng/μl of a construct expressing a sgRNA targeting GCTTCAAAAATTTCTCCAGG (derived from p46169). After isolation, PCR genotyping and confirmation by Sanger sequencing, mutants were outcrossed four times against the wild type.

Creation of endogenous tags and a *tebp-1; pot-2* double mutant via CRISPR-Cas9-mediated genome editing. Protospacer sequences were chosen using CRISPOR (<http://crispor.tefor.net>)⁷⁷, cloned in pRFK2411 (plasmid expressing Cas9 + sgRNA(F+E));⁷⁸ derived from pDD162) or pRFK2412 (plasmid expressing sgRNA(F+E)⁷⁸ with Cas9 deleted; derived from pRFK2411) via site-directed, ligase-independent mutagenesis (SLIM)^{79,80}. pDD162 (Pef3::Cas9 + Empty sgRNA) was a gift from Bob Goldstein (Addgene plasmid # 47549; <http://n2t.net/addgene:47549>; RRID:Addgene_47549)⁸¹. All plasmids were purified using NucleoSpin® Plasmid from Macherey-Nagel, eluted in sterile water and confirmed by enzymatic digestion and sequencing. All Cas9 nuclease induced double-strand breaks (DSBs) were within 20 bp distance to the desired editing site. All CRISPR-Cas9 genome editing was performed using either *dpy-10(cn64)* or *unc-58(e665)* co-conversion strategies⁸². Single-stranded oligodeoxynucleotides (ssODN, 4 nmole standard desalted Ultramer™ DNA oligo from IDT) and PCR products (purified using QIAquick® PCR Purification Kit from QIAGEN) served as donor templates for small (3xFLAG epitope tag, protospacer sequences) and big (GFP tag) insertions, respectively. The *gfp* coding sequence including three introns and flanking homology regions was amplified from pDD282, which was a gift from Bob Goldstein (Addgene plasmid # 66823; <http://n2t.net/addgene:66823>; RRID: Addgene_66823)⁸³. All donor templates contained ~35 bp homology regions^{84,85}. Plasmid vectors, ssODN and PCR products were diluted in sterile water and injected at a final concentration of 30–50 ng/μl, 500–1000 nM and 300 ng/μl, respectively. For GFP insertions, the protospacer sequence used for the *dpy-10* co-conversion was transplanted to the editing site to generate d10-entry strains⁸⁶, which in turn served as reference strains for further injections. DNA mixes were injected in both gonad arms of 10–25 1-day-old adult hermaphrodites maintained at 20 °C. Co-converted F1 progeny were screened for insertions by PCR. Successful editing events were confirmed by Sanger sequencing. All generated mutant strains were outcrossed at least two times prior to any further cross or analysis. CRISPR-Cas9 genome editing reagents and DNA injection mixes are listed in Supplementary Data file 5. The *pgl-1::mTagRfp-T* is described elsewhere^{49,50}.

Creation of transgenic worms using MosSCI. A *TEBP-2::GFP* fusion transgene was produced as previously described⁸⁷, and as indicated in www.wormbuilder.org. Animals of the strain EG6699 were injected, in order to get insertions in locus *ttT5605* on LGII. The injection mix contained all the injection constructs listed in www.wormbuilder.org, using the recommended concentrations, including 50 ng/μl of a repair template containing the *tebp-2::gfp* sequence. Selection was performed as recommended in www.wormbuilder.org⁷⁶.

Extraction of genomic DNA from *C. elegans*. Mixed-staged animals were washed off plates with M9 and washed two to three more times in M9. Next, worms were resuspended in Worm Lysis buffer (WLB: 0.2 M NaCl, 0.1 M Tris/HCl pH 8.5, 50 mM EDTA, 0.5% SDS) and aliquoted in 250 μl samples. For genomic DNA extraction the aliquots were brought to a final volume of 500 μl with WLB and Proteinase K (30 μg/ml). To lyse the worms, the samples were incubated at 65 °C at 1400 rpm for > 2 h until all carcasses were dissolved. The samples were then centrifuged at 21,000 x g for 5 min to pellet debris and the supernatant was transferred to a fresh tube. Afterwards, 500 μl of Phenol:Chloroform:

Isoamylalcohol were added, the samples shaken vigorously for 30 s and spun down at 16,000 \times g for 5 min. Additionally, 500 μ l of chloroform were added to the samples and again shaken vigorously for 30 sec and spun at 16,000 \times g for 5 min. The aqueous phase of the samples was transferred to fresh 2 ml reaction tubes and 50 μ g RNase A were added to digest the RNA. The tubes were inverted once and incubated at 37 °C for > 1 h. After RNA digestion the samples were again purified by phenol:chloroform:isoamylalcohol and chloroform addition (as before). The aqueous phase was transferred to fresh tubes and the DNA was precipitated with 350 μ l isopropanol for > 15 min at -80 °C. To pellet the DNA, the samples were centrifuged at 21,000 \times g for 20 min at 4 °C. The supernatant was carefully removed and the DNA pellet washed once with 1 ml of ice-cold 70% ethanol and spun at 21,000 \times g for 5 min at 4 °C. Washing was repeated if the samples still smelled of phenol. After washing the supernatant was completely removed, the pellet air dried for ca. 10 min, and resuspended in 20 μ l H₂O. To fully resuspend the DNA, the samples were kept at 4 °C overnight and mixed again the next day.

Telomere Southern blot. For denatured telomere Southern blot 15 μ g of *C. elegans* genomic DNA were digested in 80 μ l total volume with 40 U HinfI (New England Biolabs, #R0155) and RsaI (New England Biolabs, #R0167), respectively. The digestion was incubated at 37 °C overnight and the next day additional 10 U of each enzyme were added and the samples incubated 1–2 h further. Afterwards the samples were evaporated in a Concentrator Plus at 45 °C to end up with a volume of 20–30 μ l and supplemented with 2x DNA loading dye. A 0.6% agarose gel was prepared (with 1x TBE and 16 μ l SYBR Safe DNA stain, Thermo Fischer Scientific, #S33102) and the samples loaded after boiling at 95 °C for 10 min. The GeneRuler 1 kb (Thermo Scientific, #SM0312), as well as the 1 kb extended markers (New England Biolabs, #N3239) were used. The samples were secured in the gel by running it at 100 V for 20–30 min then the voltage was set to 60 V for a run overnight (16–19 h). With a crosslinker set to 1 min crosslinking time, the DNA was broken and the gel afterwards equilibrated in transfer buffer (0.6 M NaCl, 0.4 M NaOH) for at least 20 min. After equilibration, an upward alkaline transfer was set up with whatman paper and a positively charged nylon membrane (Byodine B membrane; Pall, #60207), all equilibrated in transfer buffer. The transfer was set up overnight. Following blotting, the membrane was fixed by incubation in 0.4 M NaOH for 15 min with slight agitation and neutralized with two washes in 2x SSC for 5 min each. To keep hydrated the membrane was sealed in cling film with 2x SSC until hybridization.

The membrane was pre-hybridized in a glass hybridization tube with 20 ml hybridization buffer (3.3x SSC, 0.1% SDS, 1 mg/ml Skim Milk powder) for at least 1 h at 42 °C rotating in a hybridization oven. The oligonucleotide used for detection was a TTAGGC reverse complement triple repeat (GCCTAA)₃. The probe was radioactively labeled with 3 μ l 32P-[γ]-ATP by a polynucleotide Kinase reaction and cleaned up using a MicroSpin Sephadex G-50 column (GE Healthcare, #GE27-5330-01). The labeled oligonucleotide was denatured at 95 °C for 10 min and mixed with 20 ml fresh hybridization buffer. This mix was added to the membrane after removing the previous buffer and incubated for 3.5 days rotating at 42 °C.

After hybridization the membrane was washed by first rinsing it twice with Wash Buffer 1 (2x SSC, 0.1% SDS), then incubating it twice for 5 min in 20 ml Wash Buffer 1. For the last wash, the membrane was incubated for 2 min in Wash Buffer 2 (0.2x SSC, 0.1% SDS), then rinsed in 2x SSC to re-equilibrate the salt concentration. The membrane was dried on a Whatman paper for 3 h, sealed in cling film and exposed to a phosphorimager screen for 3 days. The screen was read out with the Typhoon Scanner with the settings 1000 V PMT and 200 μ m pixel size. Contrast and brightness of the resulting tif-file were optimized using Fiji.

Microscopy

Co-localization microscopy. Strains carrying TEBP-1::GFP or TEBP-2::GFP were crossed with strain YA1197 expressing POT-1::mCherry. Adult animals were washed in M9 buffer, immobilized in M9 buffer supplemented with 40 mM sodium azide and mounted on freshly made 2% agarose pads. For imaging embryos, adult hermaphrodites were washed and dissected in M9 buffer before mounting. Animals were immediately imaged using a TCS SP5 Leica confocal microscope equipped with a HCX PL APO 63x water objective (NA 1.2), Leica hybrid detectors (HyD), and the acquisition software Leica LAS AF. Deconvolution was performed using Huygens Remote Manager and images were further processed using Fiji.

PGL-1 fluorescence microscopy. For imaging PGL-1::mTagRFP-T in animals of each category of germline morphology, adult worms were picked to a droplet of M9 to remove OP50 bacteria, then transferred to a drop of M9 buffer supplemented with 40 mM sodium azide in M9 for immobilization on a 2% agarose pad. Animals were immediately imaged with a Leica AF7000 widefield microscope using a 20x objective (NA 0.4) and red fluorescence filters (N3), as well as TL-DIC (acquisition software: Leica LAS X, camera: Hamamatsu, Orca Flash 4.0 V2). Images were processed using Fiji (brightness changes applied only in DIC channel for better visualization).

Quantitative FISH (qFISH). For telomere length determination, fluorescence in situ hybridization (FISH) was utilized in a quantitative manner⁸⁸. The staining protocol was optimized after the work of Seo and Lee⁸⁹. Per strain, 100 gravid adults were

picked to an unseeded small NGM plate to remove the majority of OP50 bacteria. From there, worms were picked to a 5 μ l drop of Egg buffer (25 mM HEPES/KOH pH 7.4, 118 mM NaCl, 48 mM KCl, 2 mM EDTA, 0.5 mM EGTA, 1% Tween-20) on a cover slip and dissected using 20 gauge needles (Sterican, Roth #C718.1) to release embryos and gonads. The samples were fixed by adding 5 μ l of 2% Formaldehyde solution and incubating for 5 min. To remove the Formaldehyde solution, samples were washed on the cover slip by adding and removing Egg buffer carefully by pipetting. For permeabilization of the cuticle, the worms were afterwards treated by freeze cracking⁹⁰. The cover slips were put on a Poly-lysine coated slide (Sigma Aldrich, #P0425) and the slides transferred to an aluminum block on dry ice for freezing. After 15 min freezing on the aluminum block, the cover slips were removed and the slides immersed first in ice-cold methanol, then in ice-cold acetone for 5 min, respectively. To remove the solutions the slides were washed in 1x PBS (10 mM Na₂HPO₄, 2 mM KH₂PO₄, 137 mM NaCl, 2.7 mM KCl) for 15 min. For additional permeabilization the samples were incubated in permeabilization buffer (20 mM Tris/HCl pH 7.5, 50 mM NaCl, 3 mM MgCl₂, 300 mM Sucrose, 0.5% Triton X-100) at 37 °C for 30 min followed by a wash in 1x PBS for 5 min at room temperature. To prevent unspecific binding of the FISH probe, the samples were treated with 20 μ l RNase A solution (1x PBS, 0.1% Tween-20, 10 μ g/ml RNase A) at 37 °C for 1 h in a humid chamber. Afterwards the slides were washed in 1x PBS-T (1x PBS, 0.1% Tween-20) for 10 min at room temperature and dehydrated by successive 3 min washes in 70%, 85 and 100% ethanol and air dried. For pre-hybridization 50 μ l of hybridization solution (3X SSC, 50% Formamide, 10% (w/v) Dextran-Sulfate, 50 μ g/ml Heparin, 100 μ g/ml yeast tRNA, 100 μ g/ml sheared salmon sperm DNA) were added to the sample and the slides incubated in a humid chamber for 1 h at 37 °C. The FISH probe (PNA-FISH TTAGGC telomeric probe, Panagene, resuspended to 100 μ M, fluorophore: Alexa-555) was prepared as a 1:500 dilution in hybridization solution and denatured for 5 min at 70 °C. After pre-hybridization, the solution on the slides was removed as much as possible by pipetting and 20 μ l of FISH probe were added, then covered by a cover slip. For hybridization of the probe the slides were denatured on a heat block prepared with wet paper towels for humidity at 80 °C for 3 min and transferred to a humid chamber for incubation overnight at 37 °C. The next day the slides were washed twice in 1x PBS-T for 5 min to remove the probe. To fixate the staining, the samples were incubated in hybridization wash solution (2X SSC, 50% Formamide) for 30 min at 37 °C. As a last step the slides were washed in 1x PBS-T twice for 15 min at room temperature and mounted by adding 10–20 μ l Vectashield mounting medium containing DAPI (Vector laboratories, #H-1200-10). The pictures were taken with a Leica TCS SP5 confocal microscope (objective: CX PL APO CS 63x oil NA: 1.4, pinhole 60.05 μ m, 2x zoom, PMT detectors, acquisition software Leica LAS AF). The images stacks were composed by a sequence of pictures acquired every 0.5 μ m on the z-axis. The laser and gain settings were adjusted according to the sample with the lowest FISH intensity. For analysis, images were opened in Image J/Fiji and the channels split into the DAPI and red channel. A mask of the image was created to infer the volume of the imaged object. The threshold function of the software was used with activated plugins for identification of round objects (Otsu). After setting the threshold for the image in the histogram settings, the z-stack was converted to a binary mask and using the 3D OC Options menu volume, mean gray values and integrated density of the FISH foci were calculated. Additionally, the 3D Object counter menu was used and the filters set to a minimum of 2. The values obtained by this analysis were averaged over several images of either germlines or embryos of the same strain and used for quantitative comparison of telomere length. For comparison, all values obtained for the mutant strains were scaled relative to the average of the wild type values. The barplots were created using R with standard and publicly available scripts (RCOLORBrewer-v 1.1-2, ggpubr-v 4.0, plyr-v 1.8.6, viridis-v 0.5.1, viridisLite-v 0.3.0, ggforce-v 0.3.2, ggsignif-v 0.6.0, dplyr-v 1.0.2, ggplot2-v 3.3.3, readr-v 1.4.0).

Yeast two-hybrid assay. Yeast two-hybrid assays were conducted in the yeast strain PJ69-4a as described before^{91,92}. The respective Gal4 activation and DNA-binding domain plasmid pairs were co-transformed in PJ69-4a. The resulting transformants were resuspended in ddH₂O and pinned on SC Trp-Leu-, SC Trp-Leu-His-, and SC Trp-Leu-His-Ade- plates. For Fig. 6a an additional round of plasmid transformation was performed, as a biological duplicate, and the results were identical. Colonies were imaged with a ChemiDoc XRS+ system (BioRad, Software: Image Lab 5.2.1) for Fig. 6a and Supplementary Fig. S6g, and scanned with an Epson Scanner (Perfection V700 Photo, Software version 3.81) for Fig. 6f–h and supplementary Fig. 6i.

Size-exclusion chromatography. Size-exclusion chromatography was performed as previously described^{76,92}. For the first run (Supplementary Fig. 5a) two embryo samples were prepared and combined. Using a centrifugal filter with a 10 kDa cutoff (Merck, Amicon Ultra 0.5 ml 10 K, #UFC5010) the sample was concentrated to a final volume of 550 μ l. Between 3.6 and 3.8 mg of total extract was separated on a Superose 6 10/300 GL column (GE Healthcare, 17517201) operated on a NGC Quest System (Bio-Rad) using lysis buffer without Triton X-100 as running buffer (25 mM Tris/HCl pH 7.5, 150 mM NaCl, 1.5 mM MgCl₂, 1 mM DTT, protease inhibitors). Five-hundred microliter fractions were collected according to the scheme in Supplementary table 2. Selected fractions were concentrated to 30 μ l using 10 kDa cutoff centrifugal filters (Merck, Amicon Ultra 0.5 ML 10 K,

#UFC5010). The samples were supplemented with 4x LDS (NuPAGE) and 100 mM DTT to a final volume of around 40 µl and boiled at 95 °C for 10 min. After spinning down, a part of each sample was run on a 4–15% Criterion TGX Stain-Free Protein Gel (26 wells, Bio-Rad, #5678085) in 1x SDS running buffer at 200 V for 32 min. Transfer of proteins to a nitrocellulose membrane (Bio-Rad, #1620112) was performed using the Trans-Blot Turbo Transfer System (Bio-Rad). Following the transfer, western blot was performed as described above. For the second run (Fig. 5a, b), four embryo extracts were prepared, combined and concentrated, as above, to 1 ml. Then half of the sample was treated with Sm nuclease for 30 min at 4 °C, prior to size-exclusion chromatography, while the other half was not.

Phylogenetic and synteny analysis. The protein sequences of *C. elegans* TEBP-1 and TEBP-2 were extracted from Wormbase (WS275). These sequences were used separately as queries for Wormbase BLASTP search in the available genomes. Orthologs of TEBP-1 and TEBP-2 were defined based on two criteria: (1) BLASTP hit had an E-value lower than 1.00e-15; and (2) reciprocal BLASTP of the hit, querying the *C. elegans* proteome, resulted in TEBP-1 and TEBP-2 as top hits. Sequences of the identified orthologs were obtained from Wormbase (WS275) and Wormbase ParaSite (WBPS14/WS271). The list of identified orthologs and BLASTP results can be found in Supplementary Data file 2 (sheet 1).

The full-length protein sequences of TEBP orthologs were used for multiple sequence alignment using MAFFT, version 7.452⁹³. Alignment was performed using default settings, including an automatic determination of best alignment strategy, which provided the L-INS-I result⁹⁴. Multiple sequence alignment can be found in Supplementary Data file 2 (sheet 2). Then, the multiple sequence alignment in fasta format was used as an input for IQ-TREE version 1.6.12⁹⁵, with branch supports obtained with ultrafast bootstrap⁹⁶. IQ-TREE was first ran to determine the best fit substitution model, which was VT+F+R3. Then, analysis was repeated with the following parameters: -redo -m VT+F+R3 -bb 10000 -o Cang_2012_03_13_00535.g11959_Cang, Cang_2012_03_13_01061.g15539.t3_Can, where -m is the best fit model, -bb is the number of ultrafast bootstrap replicates, and -o represents the defined outgroups. Output.tree file was visualized in FigTree version 1.4.4 (<http://tree.bio.ed.ac.uk/software/figtree/>). The *C. angaria* TEBP orthologs were used as outgroups, as this species is not part of the *Elegans* and *Japonica* groups, according to recent phylogenetic studies⁹⁷. To create an additional tree with the N-terminal region only, the initial multiple sequence alignment was trimmed to the 600 initial alignment positions. The alignment of this region (with similarity to the homeodomain of RAP1) was substantially more reliable, as assessed by higher GUIDANCE2 scores⁹⁸. Using this edited alignment, another tree was constructed as described above. IQ-TREE best fit model was VT+F+I+G4, parameters used: -m VT+F+I+G4 -bb 10000 -o Cang_2012_03_13_00535.g11959_Cang, Cang_2012_03_13_01061.g15539.t3_Can.

We defined local synteny across species as the maintenance of linkage in at least one of the neighboring genes upstream and downstream of the respective *tebp* gene. We used two different strategies to determine synteny. (1) Synteny was determined by navigating genome browser tracks through regions containing *tebp* orthologs, using Wormbase ParaSite (WBPS14/WS271). Currently annotated genes, adjacent to *tebp* orthologs, were selected, their predicted protein sequences were retrieved and BLASTP was performed in the *C. elegans* genome to find the corresponding ortholog. Results are summarized in Supplementary Data file 2 (sheet 3). (2) The protein sequences obtained previously by reciprocal BLASTP of TEBP-1 and TEBP-2 were used as an entry for WormBase ParaSite BioMart tool (<https://parasite.wormbase.org/biomart>). We recouped the neighboring 13 genes upstream and 13 genes downstream, and, with the resulting gene ID list, we determined a set of orthologous genes with the following series of 'Output attributes': gene stable ID, chromosome/scaffold, start (bp) and end (bp) coordinates that were to be listed in the result from ten available complete *Caenorhabditis* genomes. Subsequently, we filtered only those genes that share the same chromosome/scaffold with the *tebp* orthologous gene, finally, we evaluate if the enlarged group meets our definition of local synteny. We repeated this process taking each of the *tebp* genes in the ten species as a reference and evaluated the filtered groups for local synteny. In the specific case of *C. remanei*, WormBase ParaSite provides three different assemblies: PRJNA248909, PRJNA248911 and PRJNA53967. The latter was the only assembly where we were able to identify synteny of *tebp-1* with BioMart, although we could verify it manually for PRJNA248911. Results are summarized in Supplementary Data file 2 (sheet 4). This strategy was not applicable to *C. angaria*, as the genome of this species is not implemented in WormBase ParaSite BioMart.

Analysis of previously published RNA-seq datasets. For the expression data of the telomeric proteins during development of *C. elegans* (Supplementary Fig. 2a–c), RNAseq data was taken from a previously published dataset⁴⁷. To probe expression of the telomeric genes in spermatogenic and oogenic gonads (Supplementary Fig. 2d), previously published transcriptome data was used⁴⁸. Gene expression and genome browser tracks were plotted using Gviz⁹⁹ and GenomicFeatures¹⁰⁰ on an R framework (R Core Team 2018).

RNA extraction and library preparation. RNA was extracted as described⁴⁷. Synchronized young adult animals were frozen in 50–100 µl of H₂O after harvest.

After thawing, 500 µl TRIzol LS reagent (Invitrogen, # 10296010) was added and the worms were lysed with six freeze-thaw cycles (frozen in liquid nitrogen for ca. 30 s, then thawed for 2 min in a 37 °C waterbath and vortexed). Following lysis, the samples were spun down at full speed for 2 min to pellet debris. Supernatant was transferred to a fresh tube, mixed 1:1 with 100% ethanol and the mix was transferred to a column of the Direct-zol RNA MiniPrep Plus Kit (Zymo Research, #R2070). The following purification steps were done according to manufacturer's instructions, including the recommended in-column DNase I treatment for 25–40 min. RNA samples were eluted in 30–32 µl of RNase-free H₂O.

Library preparation for mRNA sequencing was performed with Illumina's TruSeq stranded mRNA LT Sample Prep Kit following Illumina's standard protocol (Part # 15031047 Rev. E). Libraries were prepared by using only ¼ of the reagents with a starting amount of 250 ng and they were amplified in ten PCR cycles. Libraries were profiled in a High Sensitivity DNA on a 2100 Bioanalyzer (Agilent technologies) and quantified using the Qubit dsDNA HS Assay Kit, in a Qubit 2.0 Fluorometer (Life technologies). Libraries were pooled in an equimolar ratio and sequenced on one NextSeq 500 Highoutput Flowcell, SR for 1 × 75 cycles plus 1 × 7 cycles for index read.

mRNA read processing and mapping. The library quality was assessed with FastQC (version 0.11.8) before alignment against the *C. elegans* genome assembly WBcel235 and a custom.GTF file, which included gene annotations from *C. elegans* (WormBase, c_elegans.PRJNA13758.WS269) and *E. coli* (EnsemblBacteria, Escherichia_coli_b_str_rel606.ASM1798v1). Alignment was performed with STAR aligner¹⁰¹ version 2.6.1b. Reads mapping to annotated features in the custom.GTF file were counted with featureCounts¹⁰² version 1.6.2 using featureCounts functionality. Counts aligning to *E. coli* were removed at this point from downstream analysis. Coverage tracks were generated with deepTools¹⁰³ version 2.27.1 and plotted using Gviz⁹⁹ on an R framework (R Core Team 2018).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets supporting the conclusions of this article are available in the ProteomeXchange Consortium via Pride repository, PXD019241; and in the SRA, BioProject PRJNA630690.

Code availability

Code is available upon request.

Received: 10 September 2020; Accepted: 30 March 2021;

Published online: 11 May 2021

References

- Jain, D. & Cooper, J. P. Telomeric strategies: means to an end. *Annu. Rev. Genet.* **44**, 243–269 (2010).
- de Lange, T. How telomeres solve the end-protection problem. *Science* **326**, 948–952 (2009).
- Doksani, Y. The response to DNA damage at telomeric repeats and its consequences for telomere function. *Genes* **10**, 318 (2019).
- Doksani, Y. & de Lange, T. The role of double-strand break repair pathways at functional and dysfunctional telomeres. *Cold Spring Harb. Perspect. Biol.* **6**, a016576 (2014).
- Lingner, J., Cooper, J. P. & Cech, T. Telomerase and DNA end replication: no longer a lagging strand problem? *Science* **269**, 1533–1535 (1995).
- Palm, W. & de Lange, T. How shelterin protects mammalian telomeres. *Annu. Rev. Genet.* **42**, 301–334 (2008).
- Soudet, J., Jolivet, P. & Teixeira, M. T. Elucidation of the DNA end-replication problem in *Saccharomyces cerevisiae*. *Mol. Cell* **53**, 954–964 (2014).
- Allsopp, R. C. et al. Telomere length predicts replicative capacity of human fibroblasts. *Proc. Natl Acad. Sci. USA* **89**, 10114–10118 (1992).
- Harley, C. B., Futcher, A. B. & Greider, C. W. Telomeres shorten during ageing of human fibroblasts. *Nature* **345**, 458–460 (1990).
- Hemann, M. T., Strong, M. A., Hao, L.-Y. & Greider, C. W. The shortest telomere, not average telomere length, is critical for cell viability and chromosome stability. *Cell* **107**, 67–77 (2001).
- Moyzis, R. K. et al. A highly conserved repetitive DNA sequence, (TTAGGG)_n, present at the telomeres of human chromosomes. *Proc. Natl Acad. Sci. USA* **85**, 6622–6626 (1988).
- de Lange, T. Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev.* **19**, 2100–2110 (2005).

13. Conomos, D., Reddel, R. R. & Pickett, H. A. NuRD–ZNF827 recruitment to telomeres creates a molecular scaffold for homologous recombination. *Nat. Struct. Mol. Biol.* **21**, 760–770 (2014).
14. Déjardin, J. & Kingston, R. E. Purification of proteins associated with specific genomic loci. *Cell* **136**, 175–186 (2009).
15. Jahn, A. et al. ZBTB48 is both a vertebrate telomere-binding protein and a transcriptional activator. *EMBO Rep.* **18**, 929–946 (2017).
16. Kappeli, D. et al. HOT1 is a mammalian direct telomere repeat-binding protein contributing to telomerase recruitment. *EMBO J.* **32**, 1681–1701 (2013).
17. Li, J. S. Z. et al. TZAP: A telomere-associated protein involved in telomere length control. *Science* **355**, 638–641 (2017).
18. Baumann, P. & Cech, T. R. Pot1, the putative telomere end-binding protein in fission yeast and humans. *Science* **292**, 1171–1175 (2001).
19. Cooper, J. P., Nimmo, E. R., Allshire, R. C. & Cech, T. R. Regulation of telomere length and function by a Myb-domain protein in fission yeast. *Nature* **385**, 744–747 (1997).
20. Miyoshi, T., Kanoh, J., Saito, M. & Ishikawa, F. Fission yeast Pot1-Tpp1 protects telomeres and regulates telomere length. *Science* **320**, 1341–1344 (2008).
21. Conrad, M. N., Wright, J. H., Wolf, A. J. & Zakian, V. A. RAP1 protein interacts with yeast telomeres in vivo: Overproduction alters telomere structure and decreases chromosome stability. *Cell* **63**, 739–750 (1990).
22. Grandin, N., Reed, S. I. & Charbonneau, M. Stn1, a new *Saccharomyces cerevisiae* protein, is implicated in telomere size regulation in association with Cdc13. *Genes Dev.* **11**, 512–527 (1997).
23. Grandin, N., Damon, C. & Charbonneau, M. Ten1 functions in telomere end protection and length regulation in association with Stn1 and Cdc13. *EMBO J.* **20**, 1173–1183 (2001).
24. Levy, D. L. & Blackburn, E. H. Counting of Rif1p and Rif2p on *Saccharomyces cerevisiae* telomeres regulates telomere length. *Mol. Cell Biol.* **24**, 10857–10867 (2004).
25. Lin, J.-J. & Zakian, V. A. The *Saccharomyces CDC13* protein is a single-strand TG1–3 telomeric DNA-binding protein in vitro that affects telomere behavior in vivo. *Proc. Natl Acad. Sci. USA* **93**, 13760–13765 (1996).
26. Moretti, P., Freeman, K., Coody, L. & Shore, D. Evidence that a complex of SIR proteins interacts with the silencer and telomere-binding protein RAP1. *Genes Dev.* **8**, 2257–2269 (1994).
27. König, P., Giraldo, R., Chapman, L. & Rhodes, D. The crystal structure of the DNA-binding domain of yeast RAP1 in complex with telomeric DNA. *Cell* **85**, 125–136 (1996).
28. Corsi, A. K., Wightman, B. & Chalfie, M. A. Transparent window into biology: a primer on *Caenorhabditis elegans*. *WormBook* 1–31 <https://doi.org/10.1895/wormbook.1.177.1> (2015).
29. Wicky, C. et al. Telomeric repeats (TTAGGC)_n are sufficient for chromosome capping function in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **93**, 8983–8988 (1996).
30. Ahmed, S. & Hodgkin, J. MRT-2 checkpoint protein is required for germline immortality and telomere replication in *C. elegans*. *Nature* **403**, 159–164 (2000).
31. Raices, M. et al. *C. elegans* telomeres contain G-strand and C-strand overhangs that are bound by distinct proteins. *Cell* **132**, 745–757 (2008).
32. Meier, B. et al. trt-1 Is the *Caenorhabditis elegans* catalytic subunit of telomerase. *PLoS Genet.* **2**, e18 (2006).
33. Cheng, C., Shtessel, L., Brady, M. M. & Ahmed, S. *Caenorhabditis elegans* POT-2 telomere protein represses a mode of alternative lengthening of telomeres with normal telomere lengths. *Proc. Natl Acad. Sci. USA* **109**, 7805–7810 (2012).
34. Lackner, D. H. & Karlseder, J. *C. elegans* survivors without telomerase. *Worm* **2**, e21073 (2013).
35. Lackner, D. H., Raices, M., Maruyama, H., Haggblom, C. & Karlseder, J. Organismal propagation in the absence of a functional telomerase pathway in *Caenorhabditis elegans*. *EMBO J.* **31**, 2024–2033 (2012).
36. Seo, B. et al. Telomere maintenance through recruitment of internal genomic regions. *Nat. Commun.* **6**, 1–10 (2015).
37. Shtessel, L. et al. *Caenorhabditis elegans* POT-1 and POT-2 repress telomere maintenance pathways. G3: genes, genomes. *Genetics* **3**, 305–313 (2013).
38. Meier, B. et al. The MRT-1 nuclease is required for DNA crosslink repair and telomerase activity in vivo in *Caenorhabditis elegans*. *EMBO J.* **28**, 3549–3563 (2009).
39. Casas-Vila, N., Scheibe, M., Freiwald, A., Kappeli, D. & Butter, F. Identification of TTAGGG-binding proteins in *Neurospora crassa*, a fungus with vertebrate-like telomere repeats. *BMC Genomics* **16**, 965 (2015).
40. Kappeli, D. et al. Phylointeractomics reconstructs functional evolution of protein binding. *Nat. Commun.* **8**, 1–9 (2017).
41. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteom.* **13**, 2513–2526 (2014).
42. Hsu, J.-L., Huang, S.-Y., Chow, N.-H. & Chen, S.-H. Stable-isotope dimethyl labeling for quantitative proteomics. *Anal. Chem.* **75**, 6843–6852 (2003).
43. Lowden, M. R., Meier, B., Lee, T. W., Hall, J. & Ahmed, S. End Joining at *Caenorhabditis elegans* telomeres. *Genetics* **180**, 741–754 (2008).
44. Zimmermann, L. et al. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
45. Lustig, A. J., Kurtz, S. & Shore, D. Involvement of the silencer and UAS binding protein RAP1 in regulation of telomere length. *Science* **250**, 549–553 (1990).
46. Li, B., Oestreich, S. & de Lange, T. Identification of human Rap1: implications for telomere evolution. *Cell* **101**, 471–483 (2000).
47. Almeida, M. V., Domingues, A. M., de, J. & Ketting, R. F. Maternal and zygotic gene regulatory effects of endogenous RNAi pathways. *PLOS Genet.* **15**, e1007784 (2019).
48. Ortiz, M. A., Noble, D., Sorokin, E. P. & Kimble, J. A New dataset of spermatogenic vs. oogenic transcriptomes in the nematode *Caenorhabditis elegans*. G3 (Bethesda, Md.) <https://doi.org/10.1534/g3.114.012351> (2014).
49. Schreier, J. et al. A membrane-associated condensate drives paternal epigenetic inheritance in *C. elegans*. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.12.10.417311> (2020).
50. Gudipati, R. K. et al. Protease-mediated processing of Argonaute proteins controls small RNA association. Preprint at *bioRxiv*. <https://doi.org/10.1101/2020.12.09.417253> (2020).
51. Kawasaki, I. et al. PGL-1, a predicted RNA-binding component of germ granules, is essential for fertility in *C. elegans*. *Cell* **94**, 635–645 (1998).
52. Strome, S. & Updike, D. Specifying and protecting germ cell fate. *Nat. Rev. Mol. Cell Biol.* **16**, 406–416 (2015).
53. Kanzaki, N. et al. Biology and genome of a newly discovered sibling species of *Caenorhabditis elegans*. *Nat. Communications* **9**, 3216 (2018).
54. Fisher, T. S. & Zakian, V. A. Ku: A multifunctional protein involved in telomere maintenance. *DNA Repair* **4**, 1215–1226 (2005).
55. Riha, K., Heacock, M. L. & Shippen, D. E. The role of the nonhomologous end-joining DNA double-strand break repair pathway in telomere biology. *Annu. Rev. Genet.* **40**, 237–277 (2006).
56. Im, S. H. & Lee, J. PLP-1 binds nematode double-stranded telomeric DNA. *Mol. Cells* **20**, 297–302 (2005).
57. Im, S. H. & Lee, J. Identification of HMG-5 as a double-stranded telomeric DNA-binding protein in the nematode *Caenorhabditis elegans*. *FEBS Lett.* **554**, 455–461 (2003).
58. Kim, S. H., Hwang, S. B., Chung, I. K. & Lee, J. Sequence-specific binding to telomeric DNA by CEH-37, a homeodomain protein in the nematode *Caenorhabditis elegans*. *J. Biol. Chem.* **278**, 28038–28044 (2003).
59. Coghlan, A. & Wolfe, K. H. Fourfold faster rate of genome rearrangement in nematodes than in drosophila. *Genome Res.* **12**, 857–867 (2002).
60. Saint-Leandre, B. & Levine, M. T. The telomere paradox: stable genome preservation with rapidly evolving proteins. *Trends Genet.* **36**, 232–242 (2020).
61. de Lange, T. Shelterin-mediated telomere protection. *Annu. Rev. Genet.* **52**, 223–247 (2018).
62. He, H. et al. POT1b protects telomeres from end-to-end chromosomal fusions and aberrant homologous recombination. *EMBO J.* **25**, 5180–5190 (2006).
63. van Steensel, B., Smogorzewska, A. & de Lange, T. TRF2 protects human telomeres from end-to-end fusions. *Cell* **92**, 401–413 (1998).
64. Long, J. et al. Telomeric TERB1–TRF1 interaction is crucial for male meiosis. *Nat. Struct. Mol. Biol.* **24**, 1073–1080 (2017).
65. Shibuya, H. et al. MAJIN links telomeric DNA to the nuclear membrane by exchanging telomere cap. *Cell* **163**, 1252–1266 (2015).
66. Wang, Y. et al. The meiotic TERB1–TERB2–MAJIN complex tethers telomeres to the nuclear envelope. *Nat. Commun.* **10**, 1–19 (2019).
67. Ferreira, H. C., Towbin, B. D., Jegou, T. & Gasser, S. M. The shelterin protein POT-1 anchors *Caenorhabditis elegans* telomeres through SUN-1 at the nuclear periphery. *J. Cell Biol.* **203**, 727–735 (2013).
68. de Albuquerque, B. F. M. et al. PID-1 is a novel factor that operates during 21U-RNA biogenesis in *Caenorhabditis elegans*. *Genes Dev.* **28**, 683–688 (2014).
69. Shevchenko, A., Tomas, H., Havli, J., Olsen, J. V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1**, 2856–2860 (2007).
70. Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A. J. R. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat. Protoc.* **4**, 484–494 (2009).
71. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
72. Studier, F. W. Protein production by auto-induction in high-density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).

73. Ball, T. K., Saurugger, P. N. & Benedik, M. J. The extracellular nuclease gene of *Serratia marcescens* and its secretion from *Escherichia coli*. *Gene* **57**, 183–192 (1987).
74. Brenner, S. The genetics of *Caenorhabditis Elegans*. *Genetics* **77**, 71–94 (1974).
75. Schweinsberg, P. J. & Grant, B. D. C. elegans gene transformation by microparticle bombardment. *WormBook* 1–10 <https://doi.org/10.1895/wormbook.1.166.1> (2013).
76. Almeida, M. V. et al. GTSF-1 is required for formation of a functional RNA-dependent RNA Polymerase complex in *Caenorhabditis elegans*. *EMBO J.* **37**, e99325 (2018).
77. Haeussler, M. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* **17**, 148 (2016).
78. Chen, B. et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479–1491 (2013).
79. Chiu, J., March, P. E., Lee, R. & Tillett, D. Site-directed, Ligase-independent mutagenesis (SLIM): a single-tube methodology approaching 100% efficiency in 4 h. *Nucleic Acids Res.* **32**, e174–e174 (2004).
80. Chiu, J., Tillett, D., Dawes, I. W. & March, P. E. Site-directed, ligase-independent mutagenesis (SLIM) for highly efficient mutagenesis of plasmids greater than 8kb. *J. Microbiological Methods* **73**, 195–198 (2008).
81. Dickinson, D. J., Ward, J. D., Reiner, D. J. & Goldstein, B. Engineering the *Caenorhabditis elegans* genome using Cas9-triggered homologous recombination. *Nat. Methods* **10**, 1028–1034 (2013).
82. Arribere, J. A. et al. Efficient marker-free recovery of custom genetic modifications with CRISPR/Cas9 in *Caenorhabditis elegans*. *Genetics* **198**, 837–846 (2014).
83. Dickinson, D. J., Pani, A. M., Heppert, J. K., Higgins, C. D. & Goldstein, B. Streamlined genome engineering with a self-excising drug selection cassette. *Genetics* **200**, 1035–1049 (2015).
84. Paix, A. et al. Scalable and versatile genome editing using linear DNAs with microhomology to Cas9 sites in *Caenorhabditis elegans*. *Genetics* **198**, 1347–1356 (2014).
85. Paix, A., Schmidt, H. & Seydoux, G. Cas9-assisted recombineering in *C. elegans*: genome editing using in vivo assembly of linear DNAs. *Nucleic Acids Res.* **44**, e128 (2016).
86. Mouridi, S. E. et al. Reliable CRISPR/Cas9 genome engineering in *Caenorhabditis elegans* using a single efficient sgRNA and an easily recognizable phenotype. *G3* **7**, 1429–1437 (2017).
87. Frøkjær-Jensen, C. et al. Single-copy insertion of transgenes in *Caenorhabditis elegans*. *Nat. Genet.* **40**, 1375–1383 (2008).
88. Lansdorp, P. M. et al. Heterogeneity in telomere length of human chromosomes. *Hum. Mol. Genet.* **5**, 685–691 (1996).
89. Seo, B. & Lee, J. Observation and quantification of telomere and repetitive sequences using fluorescence in situ hybridization (FISH) with PNA probes in *Caenorhabditis elegans*. *JoVE (Journal of Visualized Experiments)* e54224 <https://doi.org/10.3791/54224> (2016).
90. Duerr, J. S. Antibody staining in *C. Elegans* using ‘freeze-cracking’. *JoVE (Journal of Visualized Experiments)* e50664 <https://doi.org/10.3791/50664> (2013).
91. James, P., Halladay, J. & Craig, E. A. Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. *Genetics* **144**, 1425–1436 (1996).
92. Rodrigues, R. J. C. et al. PETISCO is a novel protein complex required for 21U RNA biogenesis and embryonic viability. *Genes Dev.* **33**, 857–870 (2019).
93. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160–1166 (2019).
94. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
95. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
96. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
97. Stevens, L. et al. Comparative genomics of 10 new *Caenorhabditis* species. *Evolution Lett.* **3**, 217–236 (2019).
98. Sela, I., Ashkenazy, H., Katoh, K. & Pupko, T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* **43**, W7–W14 (2015).
99. Hahne, F. & Ivanek, R. in *Statistical Genomics: Methods and Protocols* (eds. Mathé, E. & Davis, S.) 335–351 (Springer, 2016).
100. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLOS Computational Biol.* **9**, e1003118 (2013).
101. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
102. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
103. Ramirez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).

Acknowledgements

We thank the members of the Butter and Ketting laboratories for helpful discussion and Brian Luke for critical reading of the manuscript. Franziska Roth of the Butter laboratory, Bruno de Albuquerque and Svenja Hellmann of the Ketting laboratory, Laura Tomini of the Ulrich laboratory, as well as Anja Freiwald and Mario Dejung of the Proteomics core facility provided critical technical assistance. The authors thank Shawn Ahmed, the *Caenorhabditis* Genetics Center (supported by NIH Office of Research Infrastructure Programs P40 OD010440), and the National Bioresource Project for the Experimental Animal *C. elegans* (Shohei Mitani) for kindly providing *C. elegans* strains used in this study. Assistance by the following IMB core facilities is gratefully acknowledged: Media Lab, Microscopy Core Facility, Genomics Core Facility, and to Martin Möckel of the Protein Production Core Facility. This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—407023052/GRK2526/1 and Project-ID 393547839—SFB 1361. D.K. was supported by the National Research Foundation Singapore and the Singapore Ministry of Education under its Research Centres of Excellence initiative.

Author contributions

Conceptualization, S.D., M.V.A., D.K., R.F.K., and F.B.; investigation, S.D., M.V.A., E.N., J.S., N.V., C.R.; formal analysis, S.D., M.V.A., E.N., N.V., A.F.-S., A.C.-N., and F.B.; visualization, S.D., M.V.A., E.N., J.S., A.F.-S., and F.B.; writing—original draft, S.D., M.V.A., and F.B.; writing—review & editing, all authors contributed; supervision, H.D.U., R.F.K., and F.B.; project administration, F.B.; funding acquisition, H.D.U., R.F.K., F.B.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-22861-2>.

Correspondence and requests for materials should be addressed to F.B.

Peer review information *Nature Communications* thanks Jerome Dejaridin, Jan Karlseder and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021, corrected publication 2021

4.10.5 Article I: Supplementary information

SUPPLEMENTARY INFORMATION

The double-stranded DNA-binding proteins TEBP-1 and TEBP-2 form a telomeric complex with POT-1

Sabrina Dietz^{1,5}, Miguel Vasconcelos Almeida^{1,2,5}, Emily Nischwitz¹, Jan Schreier¹, Nikenza Viceconte¹, Albert Fradera-Sola¹, Christian Renz¹, Alejandro Ceron-Noriega¹, Helle D. Ulrich¹, Dennis Kappei^{3,4}, René F. Ketting¹, Falk Butter^{1,*}

¹Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz, Germany

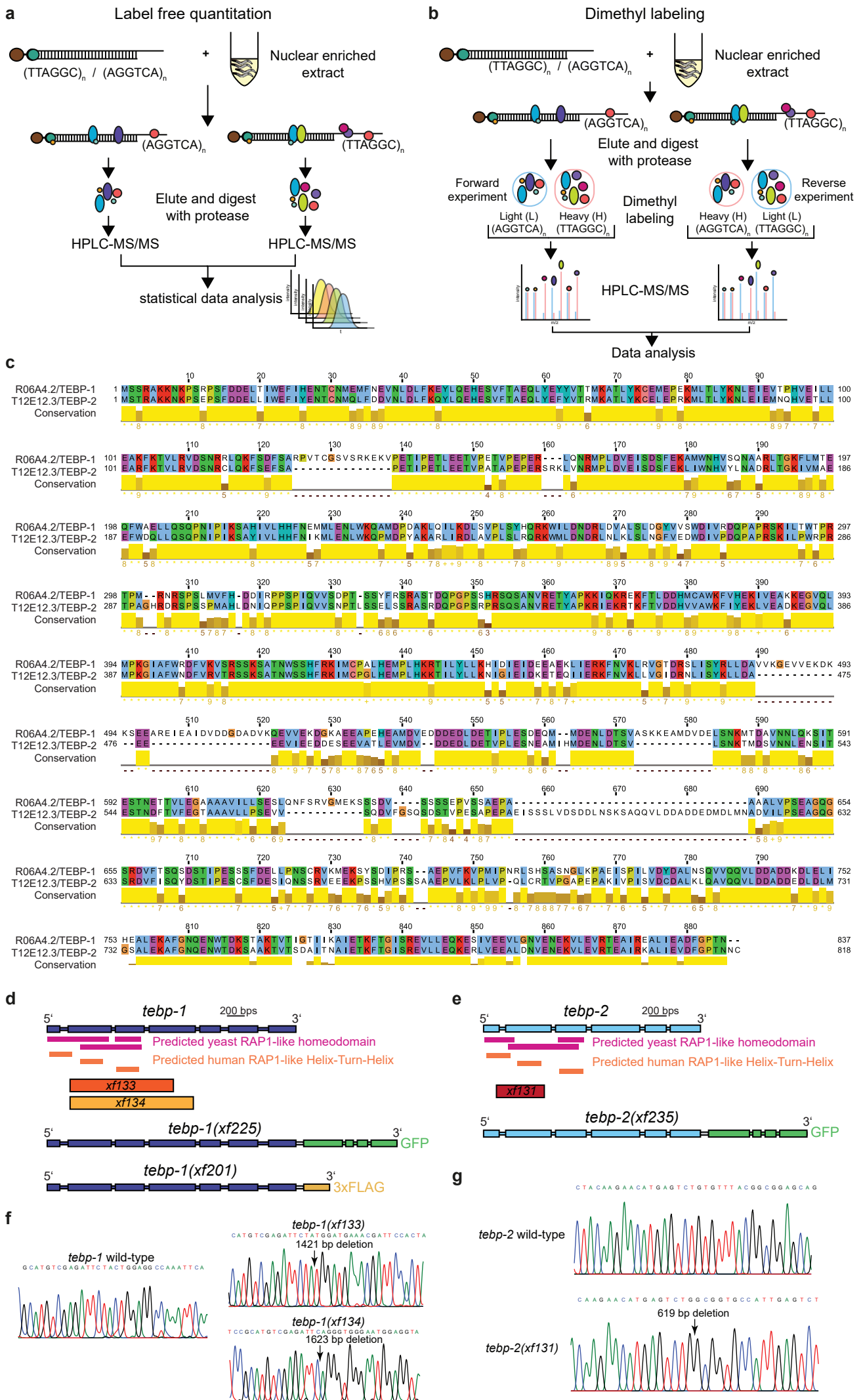
²Present address: Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QN, UK; and Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK.

³Cancer Science Institute of Singapore, National University of Singapore, 117599, Singapore

⁴Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 117596 Singapore

⁵These two authors contributed equally to this work

*Correspondence: f.butter@imb-mainz.de, +49 6131-39-21570



Supplementary Fig. 1

Supplementary Fig. 1. A quantitative proteomics screen for telomere binders identifies the paralogs TEBP-1 and TEBP-2.

(a) Scheme representing the label free quantitation workflow. Telomere (TTAGGC)_n, or control DNA (AGGTCA)_n baits are incubated with nuclear extract. Samples are processed and measured independently, and later compared by statistical data analysis.

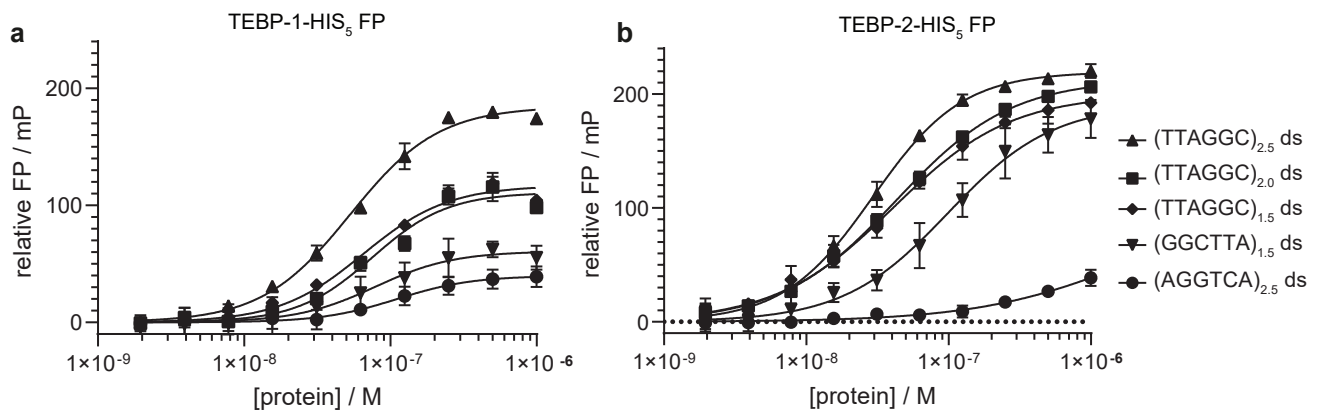
(b) Scheme representing the reductive dimethyl labeling workflow. Telomere (TTAGGC)_n, or control DNA (AGGTCA)_n baits are incubated with nuclear extract in duplicates. Per condition each peptide gets labeled with either light methyl groups (CH₃) or heavy methyl groups (CD₃). Afterwards, the heavy sample of one condition is combined with the light sample of the other condition and vice-versa to achieve a forward and a reverse experiment. Forward and reverse experiments are measured and analyzed by comparing intensities of the proteins (calculated from their peptide intensities) in the respective channel.

(c) Pairwise sequence alignment of amino acid sequences of TEBP-1 and TEBP-2 using EMBOSS Needle, visualized using Jalview, showing the high sequence similarity between the two proteins. Amino acids are color coded according to the Clustal X colour scheme: blue – amino acids A, I, L, M, F, W, C and V; red: amino acids R and K; green – amino acids N, S, Q, T; pink – amino acid C; magenta – amino acids E and D; orange – amino acid G; cyan – amino acids H, Y; yellow – amino acid P. Conservation is shown in the yellow bars beneath the sequences, brighter yellow for higher conservation. Amino acid positions are indicated.

(d) Scheme of the *tebp-1* genomic locus. Below are indicated the positions with similarity to the homeodomain of human and yeast RAP1, as predicted by HHPred (3.2.0), deletions made by CRISPR-Cas9 genome editing (alleles *xf133* and *xf134*), as well as the locations of the tags (C-terminal GFP and 3xFLAG), also inserted by CRISPRCas9 genome editing.

(e) As in (d) but for the *tebp-2* locus.

(f-g) Chromatograms of Sanger sequencing of *tebp-1* and *tebp-2* deletion alleles compared to WT. Deletion sites are indicated with arrows. Colors indicate the different DNA bases: black – G; blue – C; red – T; green – A.



c

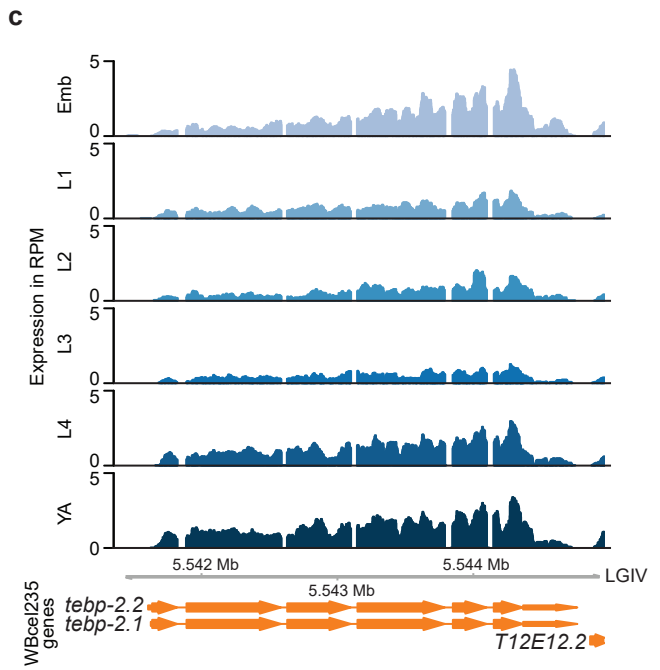
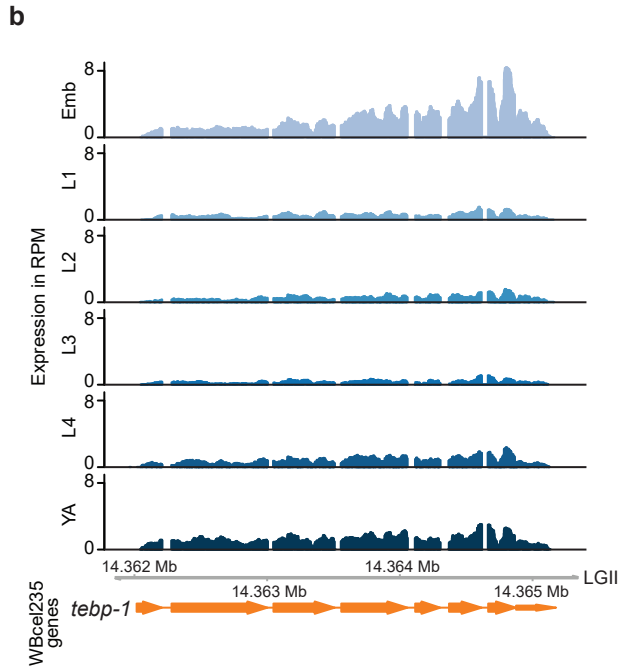
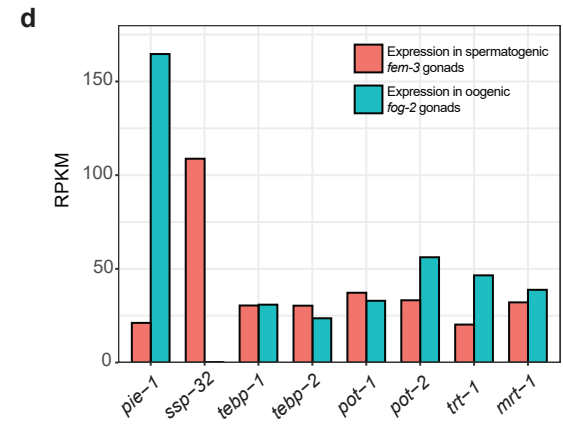
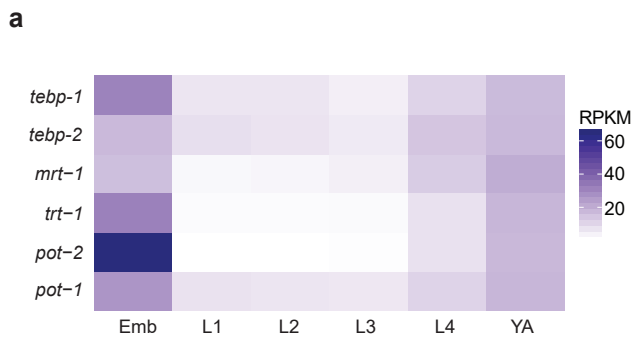
Protein	Oligo Sequence	K_d	B_{max}
TEBP-1-HIS ₅	(TTAGGC) _{2.5} ds	53.05 nM	184.5
TEBP-1-HIS ₅	(TTAGGC) _{2.0} ds	75.25 nM	111.1
TEBP-1-HIS ₅	(TTAGGC) _{1.5} ds	65.26 nM	117.0
TEBP-1-HIS ₅	(GGCTTA) _{1.5} ds	78.74 nM	60.80
TEBP-1-HIS ₅	(AGGTCA) _{2.5} ds	110.6 nM	39.88
TEBP-2-HIS ₅	(TTAGGC) _{2.5} ds	29.37 nM	219.6
TEBP-2-HIS ₅	(TTAGGC) _{2.0} ds	42.77 nM	213.2
TEBP-2-HIS ₅	(TTAGGC) _{1.5} ds	40.53 nM	199.8
TEBP-2-HIS ₅	(GGCTTA) _{1.5} ds	99.08 nM	191.3
TEBP-2-HIS ₅	(AGGTCA) _{2.5} ds	195.4 nM	104.9

Supplementary Fig. 2

Supplementary Fig. 2. Telomeric double-strand binding preferences of TEBP-1 (R06A4.2) and TEBP-2 (T12E12.3).

(a-b) Fluorescence polarization assays of 1 μ M to 2 nM purified TEBP-1-His₅ (a) and TEBP-2-His₅ (b). Proteins were incubated with 2.5x, 2.0x, 1.5x T-rich, and 1.5x G-rich double-stranded telomeric FITC-labeled oligonucleotides, as well as 2.5x double-stranded control. Error bars represent +/- the standard deviation of the mean values. Per data point n=3 technical replicates. FP, fluorescence polarization; mP, millipolarization, upward triangle: 2.5x TTAGGC double-strand, downward triangle: 2.5x TTAGGC single-strand, square: 2x TTAGGC double-strand, diamond: 1.5x TTAGGC T-rich double-strand, downward triangle: 1.5x G-rich GGCTAA double-strand, circle: 2.5x shuffled control AGGTCA double-strand.

(c) Overview of K_d and B_{max} values from FP experiment (a-b).



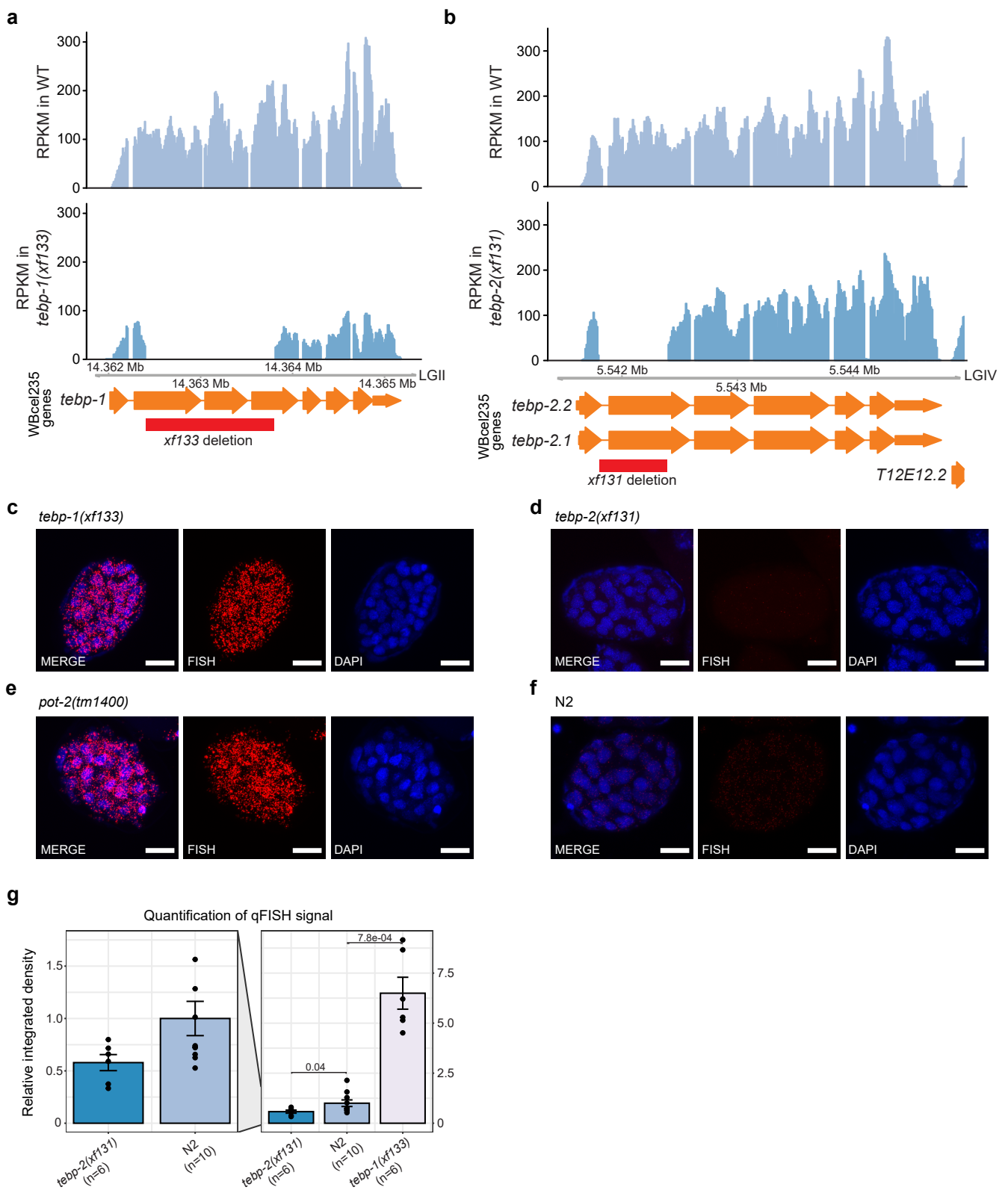
Supplementary Fig. 3

Supplementary Fig. 3. The expression profiles of *tebp-1* and *tebp-2* throughout development and in isolated gonads.

(a) Heatmap depicting mRNA expression levels, in Reads Per Kilobase Million (RPKM), of the known telomere binders *pot-1*, *pot-2*, and *mrt-1*, telomerase subunit *trt-1*, as well as *tebp-1* and *tebp-2*. Data from a previously published RNA-seq dataset⁴⁷.

(b-c) Genome browser tracks with the mRNA expression of *tebp-1* (b), and *tebp-2* (c), in reads per million (RPM), across the different life stages of *C. elegans*. Data from [47]. (a-c) Emb, embryos; L1-L4, first to fourth larval stages; YA, young adults.

(d) Expression of telomere factors in dissected *fem-3* mutant gonads (exclusively spermatogenic) and *fog-2* mutant gonads (exclusively oogenic), from previously published RNA-seq data⁴⁸. *pie-1* and *ssp-32* are genes known to be expressed in oogenesis and in spermatogenesis, respectively, according to [48].



Supplementary Fig. 4

Supplementary Fig. 4. TEBP-1 and TEBP-2 regulate telomere length in embryos.

(a-b) Genome browser tracks with the mRNA expression of *tebp-1* (a) and *tebp-2* (b), in Reads Per Kilobase Million (RPKM). RNA-seq data of wild-type, *tebp-1(xf133)*, and *tebp-2(xf131)* mutants.

(c-f) Representative maximum projection z-stacks of a qFISH assay using embryos of *C. elegans* mutant strains. The telomeres of these embryos were visualized by hybridization with a telomeric PNA-FISH-probe. Nuclei were stained with DAPI. Scale bars, 10 μ m. *tebp-1(xf133)* and *tebp-2(xf131)* were grown for approx. 98/120 generations before the experiment. N = 3 biologically independent experiments with similar results.

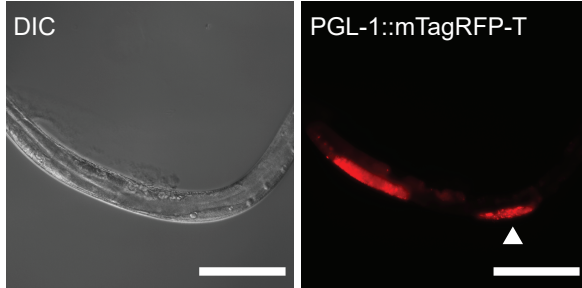
(g) Barplot depicting analysis of qFISH images of the strains in (c-f), as indicated on the x-axis. Average telomere length is indicated by arbitrary units of relative integrated density on the y-axis, with wild-type N2 set to 1. The left hand plot is a zoomed-in inset of the N2 and *tebp-2(xf131)* values. n of analyzed independent embryos per strain: *tebp-2(xf131)*: n=6, N2: n=10, *tebp-1(xf131)*: n=6. Error bars represent the standard error of the mean (SEM) and p-values were calculated using Welch's t-test.

a

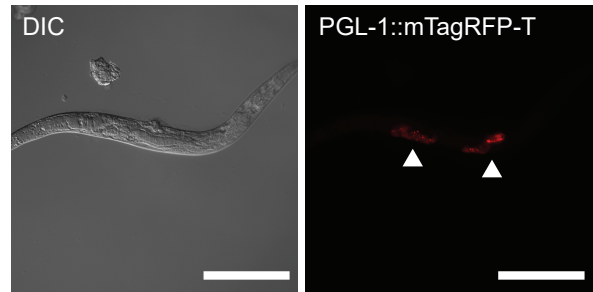
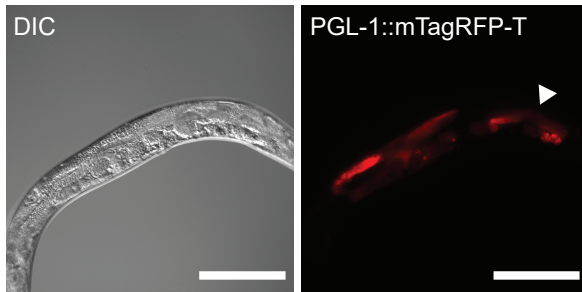
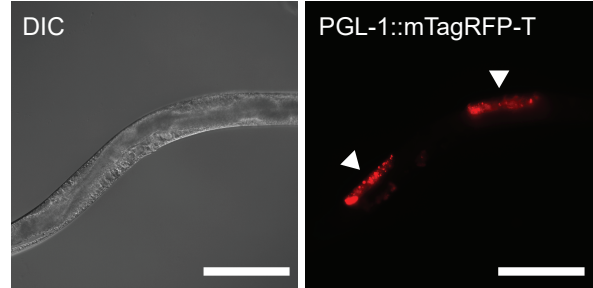
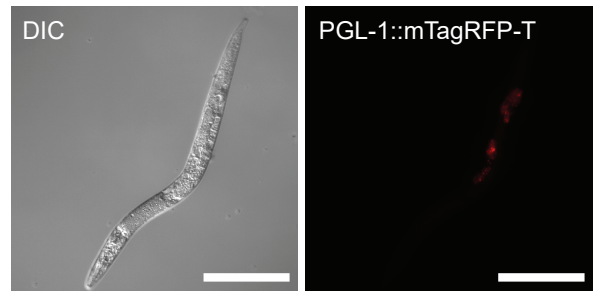
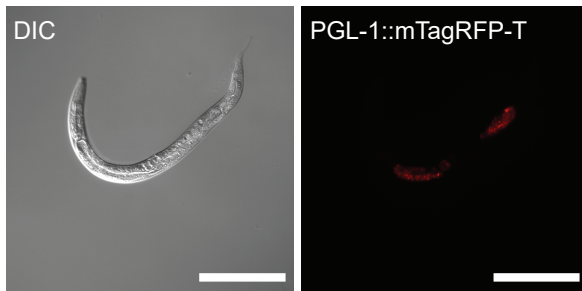
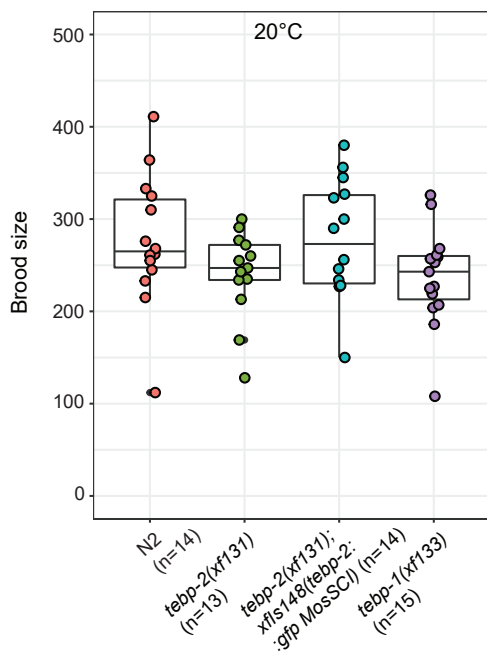
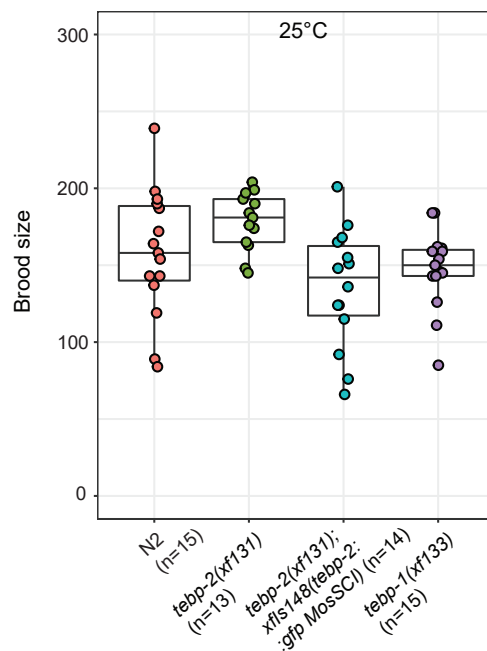
Genotype	Father	Mother	Synthetic sterile F2	Could grow double mutant homozygous line
<i>tebp-2(xf131); tebp-1(xf133)</i>	<i>tebp-2(xf131)</i>	<i>tebp-1(xf133)</i>	Yes	No
<i>tebp-1(xf133); tebp-2(xf131); xfls148(tebp-2::gfp MosSCI)</i>	<i>tebp-1(xf133)</i>	<i>tebp-2(xf131); xfls148(tebp-2::gfp MosSCI)</i>	No	Yes
<i>tebp-2(xf131); tebp-1(xf133)</i>	<i>tebp-1(xf133)</i>	<i>tebp-2(xf131); xfls148(tebp-2::gfp MosSCI)</i>	Yes	No
<i>tebp-2(xf131); pot-2(tm1400)</i>	<i>tebp-2(xf131)</i>	<i>pot-2(tm1400)</i>	No	Yes
<i>tebp-1(xf131); trt-1(ok410)</i>	<i>tebp-1(xf131)</i>	<i>trt-1(ok410)</i>	No	Yes
<i>tebp-1(xf133); mrt-1(tm1354)</i>	<i>tebp-1(xf133)</i>	<i>mrt-1(tm1354)</i>	No	Yes
<i>pot-2(tm1400); trt-1(ok410)</i>	<i>pot-2(tm1400)</i>	<i>trt-1(ok410)</i>	No	Yes
<i>tebp-2(xf131); trt-1(ok410)</i>	<i>tebp-2(xf131)</i>	<i>trt-1(ok410)</i>	No	Yes
<i>tebp-1(xf260); pot-2(tm1400)</i>	N/A. Used CRISPR-Cas9 to introduce <i>tebp-1</i> mutation due to linkage	<i>pot-2(tm1400)</i>	N/A not a cross	Yes

b

Category 2: one gonad arm atrophied



Category 3: both gonad arms atrophied

**c****d****e**

Supplementary Fig. 5

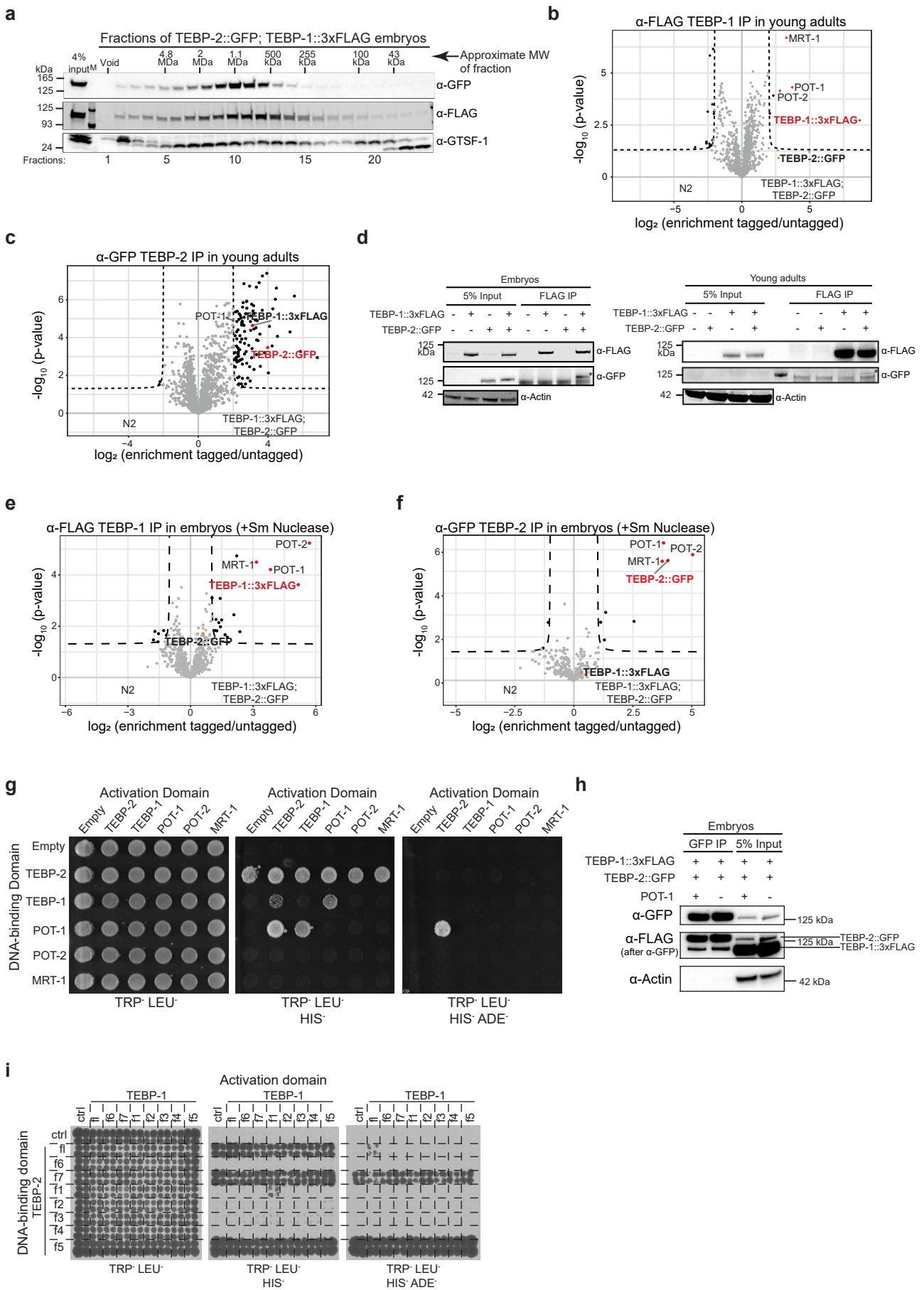
Supplementary Fig. 5. Dissecting the role of TEBP-1 and TEBP-2 in fertility.

(a) Overview of additional crosses performed to investigate distinct aspects of the synthetic sterility phenotype. For each cross, the columns indicate the genotype of the animals analyzed, the genotype of their parents, whether the animals have synthetic sterility, and if we could establish a homozygous line. The second row shows that the reciprocal cross between *tebp-1* and *tebp-2* also led to synthetic sterility. The third row shows that a *tebp-2::gfp* single-copy transgene rescues the synthetic sterility of *tebp-1; tebp-2* double mutants, while their transgene-less siblings still display synthetic sterility (fourth row). The following rows demonstrate that the synthetic sterility is specific to *tebp-1* and *tebp-2*, as it does not arise in crosses with other telomere-associated mutants.

(b) Additional representative widefield DIC and fluorescence pictures of worms with germlines of categories 2 (left panels) and 3 (right panels). Scale bars, 200 μm . Atrophied germlines are indicated with white arrowheads.

(c) Exemplary widefield DIC and fluorescence micrographs of worms showing growth defects and/or larval arrest. These animals were isolated concurrently to animals shown in (b), but did not reach adulthood. These two specific animals were offspring of *tebp-2(xf131); tebp-1(xf133) +/-*. Scale bars, 200 μm .

(d-e) Boxplots showing the brood sizes of wild-type N2, *tebp-1* or *tebp-2* single mutants, and *tebp-2(xf131); xfls148(tebp-2::gfp)*. Central horizontal lines represent the median, the bottom and top of the box represent the 25th and 75th percentile, respectively. Whiskers represent the 5th and 95th percentile, dots represent the data points used to calculate the box plot. Experiments were carried out at 20°C (d) and 25°C (e). Statistical comparisons were performed with wildtype N2, calculated with two-sided and unpaired Mann–Whitney and Wilcoxon tests. N2 vs. *tebp-2(xf131)*: 20°C p-value=0.145, 25°C p-value=0.097; N2 vs. *tebp-2(xf131); xfls148(tebp-2::gfp MosSCI)*: 20°C p-value=0.91, 25°C p-value=0.183; N2 vs. *tebp-1(xf133)*: 20°C p-value=0.052, 25°C p-value=0.41. Analyzed individuals per strain are indicated as n on the x-axis labels.



Supplementary Fig. 6

Supplementary Fig. 6. TEBP-1 and TEBP-2 interact with each other and with POT-1/MRT-1/POT-2.

(a) Western blot of the eluted fractions from size-exclusion chromatography of embryo extracts containing TEBP-1::3xFLAG and TEBP-2::GFP. The approximate molecular weight (MW) of the fractions is indicated above the blots. GTSF-1 was used as a control, as it has a known elution profile in size-exclusion chromatography⁷⁶. Information about α -GTSF-1 can be found in [76]. N = 2 biologically independent experiments with similar results.

(b-c) Volcano plots with quantitative proteomic analysis of TEBP-1::3xFLAG (b) or TEBP-2::GFP (c) IPs in young adults. IPs were performed in quadruplicates. Enriched proteins (threshold: 4-fold, p-value<0.05) are shown as black dots, enriched proteins of interest are highlighted with red or orange dots, and the baits are named in red. Background proteins are depicted as grey dots.

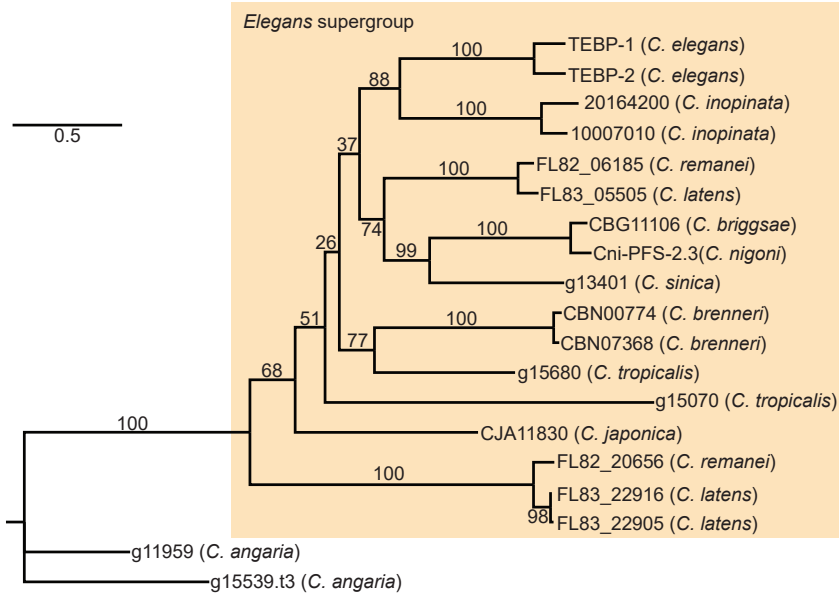
(d) Co-IP western blot experiment of TEBP-1::3xFLAG and TEBP-2::GFP similar to Fig. 5e-f, except the IPs were performed with an α -FLAG antibody. Actin was used as loading control. IPs with embryo extracts in the left panel and with young adult extracts in the right panel. N = 3 biologically independent experiments with similar results for both experiments.

(e-f) Volcano plots showing quantitative proteomic analysis of either TEBP-1::3xFLAG (e) or TEBP-2::GFP (f) IPs in embryos. IPs were performed in quadruplicates and Sm nuclease was added to remove potential DNA-dependent interactions. Enriched proteins (threshold >2-fold, p-value<0.05) are shown as black dots. Enriched proteins of interest are highlighted with red or orange dots, and baits are named in red.

(g) Orthogonal grid of the Y2H spotting containing fusion constructs of the Gal4 activating or DNA-binding domains with the full length sequence of telomere factors. Left panel shows growth control in non-restrictive medium. Protein-protein interactions allow for growth on TRP- LEU- HIS- medium (middle panel). TEBP-2 bound to the Gal4 DNA-binding domain is self-activating, precluding the determination of interactions. The strongest interactions are permissive of growth on the highly stringent TRP- LEU- HIS- ADE- medium (right panel).

(h) Co-IP western blot experiments of TEBP-1::3xFLAG and TEBP-2::GFP in the presence and absence of POT-1, where absence of POT-1 refers to the *pot-1(tm1620)* mutation. The IPs were performed with an α -GFP antibody. Actin was used as loading control. IPs were performed with 800 μ g of embryo extracts. Detection by ECL was performed sequentially, first for GFP and then for FLAG.

(i) Y2H spotting as in (g) with TEBP-1 and TEBP-2 partial constructs fused to the GAL4 activation or DNA-binding domain as in Fig. 6d,h. The full length, f7, and f5 TEBP-2 constructs fused to the Gal4 DNA-binding domain show self-activation. As in (g) the growth on the highly stringent TRP-LEU-HIS-ADE-medium (right panel) indicates the strongest interactions.



Supplementary Fig. 7

Supplementary Fig. 7. Phylogenetic analysis of the N-terminal region of TEBP-proteins.

Phylogenetic tree constructed as in Fig. 7a. The MAFFT protein alignment used for this tree comprised the first 600 alignment positions of the multiple sequence alignment in Supplementary Data 2 (sheet 2).

Values on the nodes represent bootstrapping values of 10000 replicates, set to 100.

Supplementary Table 1: List of strains used and created in this study.

Listed are all strains with their respective genotype and source.

Strain Reference	Genotype	Source
	Wild-Type N2	CGC
YA1197	<i>ypln2 [daz-1p::pot-1::mCherry::tbb-2 3'UTR + Cbr-unc-119(+)] II.</i>	A kind gift from Shawn Ahmed
tm1620	<i>pot-1(tm1620) III.</i>	National Bioresource Project for the nematode, Japan
tm1400	<i>pot-2(tm1400) II.</i>	National Bioresource Project for the nematode, Japan
YA1116	<i>mrt-1(tm1354) I.</i>	CGC
YA1059	<i>trt-1(ok410) I.</i>	CGC
EG6699	<i>ttTi5605 II; unc-119(ed3) III; oxEx1578</i>	CGC
RFK641	<i>tebp-2(xf131) IV.</i>	This study
RFK671	<i>tebp-1(xf133) II.</i>	This study
RFK672	<i>tebp-1(xf134) II.</i>	This study
RFK659	<i>TEBP-2(xfls148[tebp-2(prm)::tebp-2::GFP::tebp-2(3'UTR)]) II; unc-119(ed9) III.</i>	This study
RFK1096	<i>tebp-2(xf235[TEBP-2::GFP]) IV.</i>	This study
RFK1022	<i>tebp-1(xf225[tebp-1::GFP]) II.</i>	This study
RFK958	<i>tebp-1(xf201[tebp-1::3xFLAG]) II.</i>	This study
RFK1173	<i>tebp-2(xf235[tebp-2::GFP]) IV; tebp-1(xf201[tebp-1::3xFLAG]) II.</i>	This study
RFK1174	<i>tebp-2(xf235[tebp-2::GFP]) IV; ypln2[daz-1p::pot-1::mCherry::tbb-2 3'UTR + Cbr-unc-119(+)] II.</i>	This study
RFK1067	<i>tebp-1(xf225[tebp-1::GFP]) II; ypln2[daz-1p::pot-1::mCherry::tbb-2 3'UTR + Cbr-unc-119(+)] II.</i>	This study
RFK1086	<i>pgl-1(xf233[pgl-1::mTagRFP-T]) IV.</i>	Jan Schreier, Ketting laboratory
RFK1327	<i>tebp-2(xf131) IV; pgl-1(xf233[pgl-1::mTagRFP-T]) IV.</i>	This study
RFK1328	<i>tebp-1(xf133) II; pgl-1(xf233[pgl-1::mTagRFP-T]) IV.</i>	This study
-	<i>tebp-2(xf131) IV; pot-2(tm1400) II.</i>	This study
-	<i>tebp-1(xf133) II; mrt-1(tm1354) I.</i>	This study
RFK1334	<i>trt-1(ok410) I; tebp-1(xf133) II.</i>	This study
RFK1309	<i>tebp-1(xf260) II; pot-2(tm1400) II.</i>	This study
-	<i>trt-1(ok410) I; pot-2(tm1400) II.</i>	This study
AF16	<i>C. briggsae</i> Wild-type	CGC

Supplementary Table 2. Fractions of the gel filtration runs and correlated molecular weight.

Separation range of the used column in red, fractions covered by the marker run in green. Fractions of the 96-well column marked in bold were concentrated and used for western blot detection (Figs. 5a and S6a respectively). MW: molecular weight.

Fraction	volume [ml]	log MW	calculated MW [kDa]	96 well
A1	1,0	8,982	960063,591	a1
A2	2,0	8,727	533212,105	a2
A3	3,0	8,472	296141,997	a3
A4	4,0	8,216	164475,040	a4
A5	5,0	7,961	91348,201	a5
A6	6,0	7,705	50734,105	a6
A7	6,5	7,578	37809,419	a7
A8	7,0	7,450	28177,340	a8
A9	7,5	7,322	20999,067	a9
A10	8,0	7,195	15649,483	a10
A11	8,5	7,067	11662,724	a11
A12	9,0	6,939	8691,605	a12
A13	9,5	6,811	6477,389	b12
A14	10,0	6,684	4827,252	b11
A15	10,5	6,556	3597,493	b10
A16	11,0	6,428	2681,020	b9
A17	11,5	6,301	1998,021	b8
A18	12,0	6,173	1489,018	b7
A19	12,5	6,045	1109,686	b6
A20	13,0	5,918	826,990	b5
A21	13,5	5,790	616,311	b4
A22	14,0	5,662	459,304	b3
A23	14,5	5,534	342,295	b2
A24	15,0	5,407	255,094	b1
A25	15,5	5,279	190,108	c1
A26	16,0	5,151	141,677	c2
A27	16,5	5,024	105,584	c3
A28	17,0	4,896	78,686	c4
A29	17,5	4,768	58,641	c5
A30	18,0	4,641	43,702	c6
A31	18,5	4,513	32,569	c7
A32	19,0	4,385	24,272	c8
A33	19,5	4,257	18,088	c9
A34	20,0	4,130	13,480	c10
A35	20,5	4,002	10,046	c11
A36	21,0	3,874	7,487	c12
A37	21,5	3,747	5,580	d12
A38	22,0	3,619	4,158	d11
A39	22,5	3,491	3,099	d10
A40	23,0	3,364	2,309	d9
A41	23,5	3,236	1,721	d8
A42	24,0	3,108	1,283	d7

Superose 6 column separation range (5-5000 kDa)

covered by marker run

4.11 Article II: Nematode gene annotation by machine-learning-assisted proteotranscriptomics enables proteomewide evolutionary analysis

4.11.1 Summary

In this project, we demonstrated the effectiveness of integrating a cutting-edge proteotranscriptomics approach with machine learning quality control in obtaining high-quality gene annotations in nematodes. The approach resulted in the generation of high-quality protein-coding gene annotations for 12 species, including some species outside of the *Caenorhabditis* lineage, and led to the discovery of 2 previously unknown genes in *C. elegans*. The study also revealed the mistakes in some of the previously provided genome annotations, e.g. it identified hundreds of falsely merged genes in *P. pacificus*. Furthermore, our generic pipeline was also used to provide annotations for three species whose genomes had not been previously sequenced or assembled. An orthology analysis identifying 23,090 orthology groups across the 12 species was conducted, followed by a positive selection analysis on up to 5,400 orthologous groups and an enrichment analysis, which implied that nematode species have evolved to enhance their adaptation to their surroundings through changes in genes involved in stress response, detoxification, metabolism, reproduction, and development. This study highlights the power of the broad dataset for evolutionary analyses and offers a valuable foundation for future evolutionary investigations.

4.11.2 Zusammenfassung

In diesem Projekt haben wir demonstriert wie die Integration eines hochmodernen Proteotranskriptomik-Ansatzes mit der Qualitätskontrolle des maschinellen Lernens qualitativ hochwertige Genannotationen in Nematoden produziert. Der Ansatz führte zur Generierung hochakkuratere proteinkodierender Genannotationen für 12 Arten, darunter einige Arten außerhalb der *Caenorhabditis*-Linie, und führte zur Entdeckung von 2 zuvor unbekanntem Genen in *C. elegans*. Die Studie deckte auch die Fehler in einigen der zuvor bereitgestellten Genomannotationen auf, so identifizierte sie Hunderte von fälschlicherweise legierten Genen in *P. pacificus*. Darüber hinaus verwendeten wir unsere generische Pipeline auch, um Annotation für drei Arten bereitzustellen, deren Genome zuvor nicht sequenziert oder assembled worden waren. Es wurde eine Orthologieanalyse durchgeführt, bei der 23.090 Orthologiegruppen über die 12 Arten hinweg identifiziert wurden. Eine positive Selektionsanalyse von bis zu 5.400 Orthologengruppen und einer Anreicherungsanalyse implizierten, dass sich Nematodenarten an ihre Umgebung anpassen, indem sie die relevanten Gene im Bereich der Stressreaktion, Entgiftung, Stoffwechsel, Fortpflanzung und Entwicklung verändern und so ihre Aktivität verbessern. Die Studie unterstreicht die Leistungsfähigkeit des breiten Datensatzes für evolutionäre Analysen und bietet eine wertvolle Grundlage für zukünftige evolutionäre Untersuchungen.

4.11.3 Statement of Contribution

4.11.3 Statement of Contribution

As the first author, I performed all critical steps in this Proteotranscriptomic study. I successfully cultured the nematodes and extracted the total RNA for sequencing. I also extracted the proteome and prepared the samples for mass spectrometry measurements, including in-gel digestion and StageTip purification. I assembled the transcriptomes for 12 species both in the genome-free as well as genome-guided mode (where possible) and assessed the quality of the assemblies. I performed the annotation of the transcriptomes and developed the machine-learning approach for transcript completeness prediction. Furthermore, I conducted enrichment analysis and performed protein orthology searches and phylogenetic relation analyses. To facilitate the genome-wide positive selection analysis across the included nematode species I developed a pipeline including all major steps of the analysis. Michal Levin and I assembled and finalized all figures for the manuscript, which we wrote together with Dr. Butter.

Supervisor confirmation Falk Butter

4.11.4 Article II: Main Text

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}¹Institute 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 high-

ly 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

²These authors contributed equally to this work.³Present address: Wellcome Trust/Cancer Research UK Gurdon Institute and Department of Genetics, University of Cambridge, Cambridge CB2 1QN, UKCorresponding authors: m.levin@imb.de, f.butter@imb.deArticle published online before print. Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.277070.122>.© 2023 Ceron-Noriega et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <https://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

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 length 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

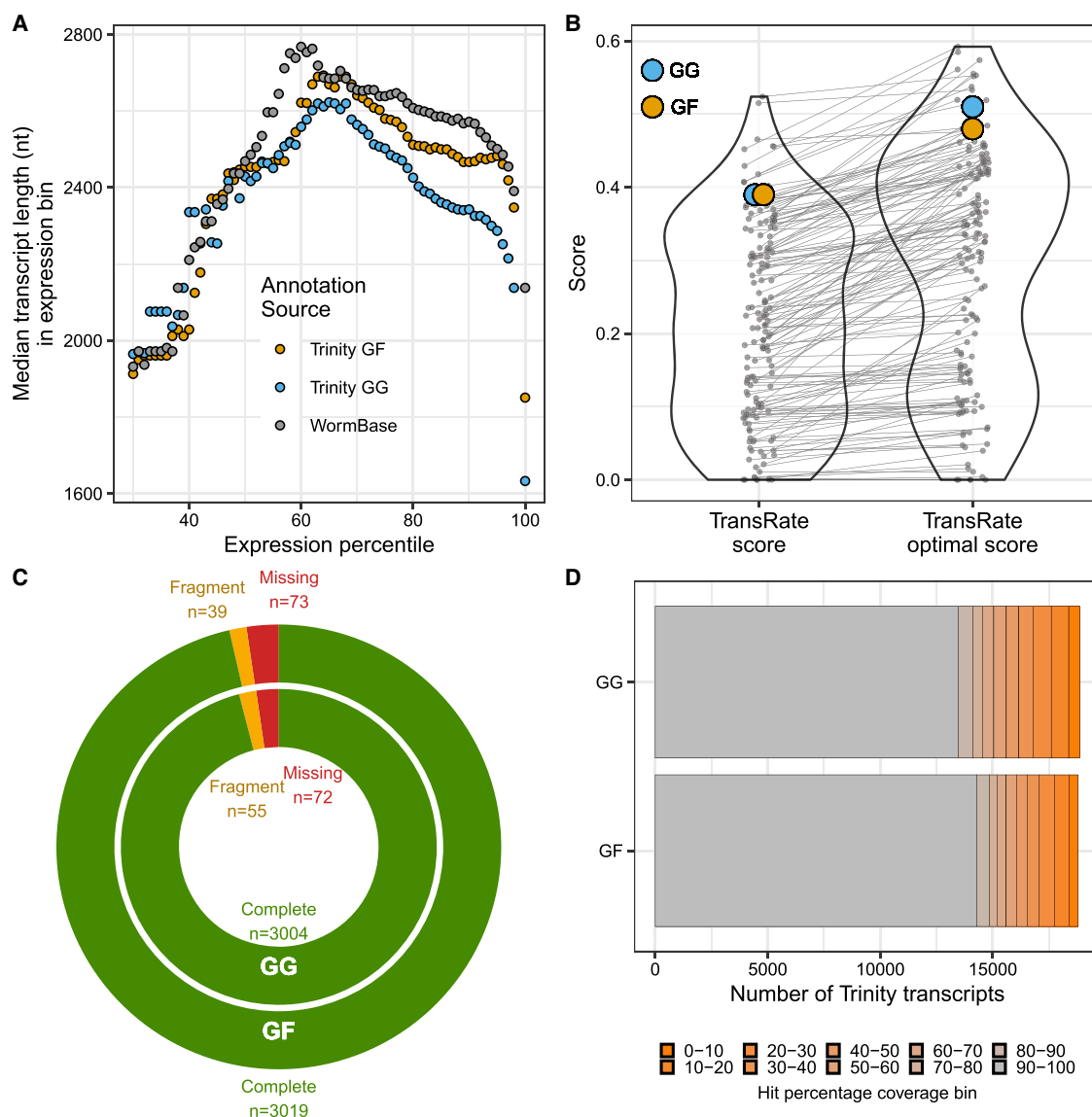


Figure 1. Benchmarking transcriptome assembly in *C. elegans*. (A) Median length of transcripts (N50) across all expression bins (ExN50) for the genome-free (GF) and genome-guided (GG) assembly compared with the WormBase annotation. (B) TransRate scores and TransRate optimal scores of the *C. elegans* 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.

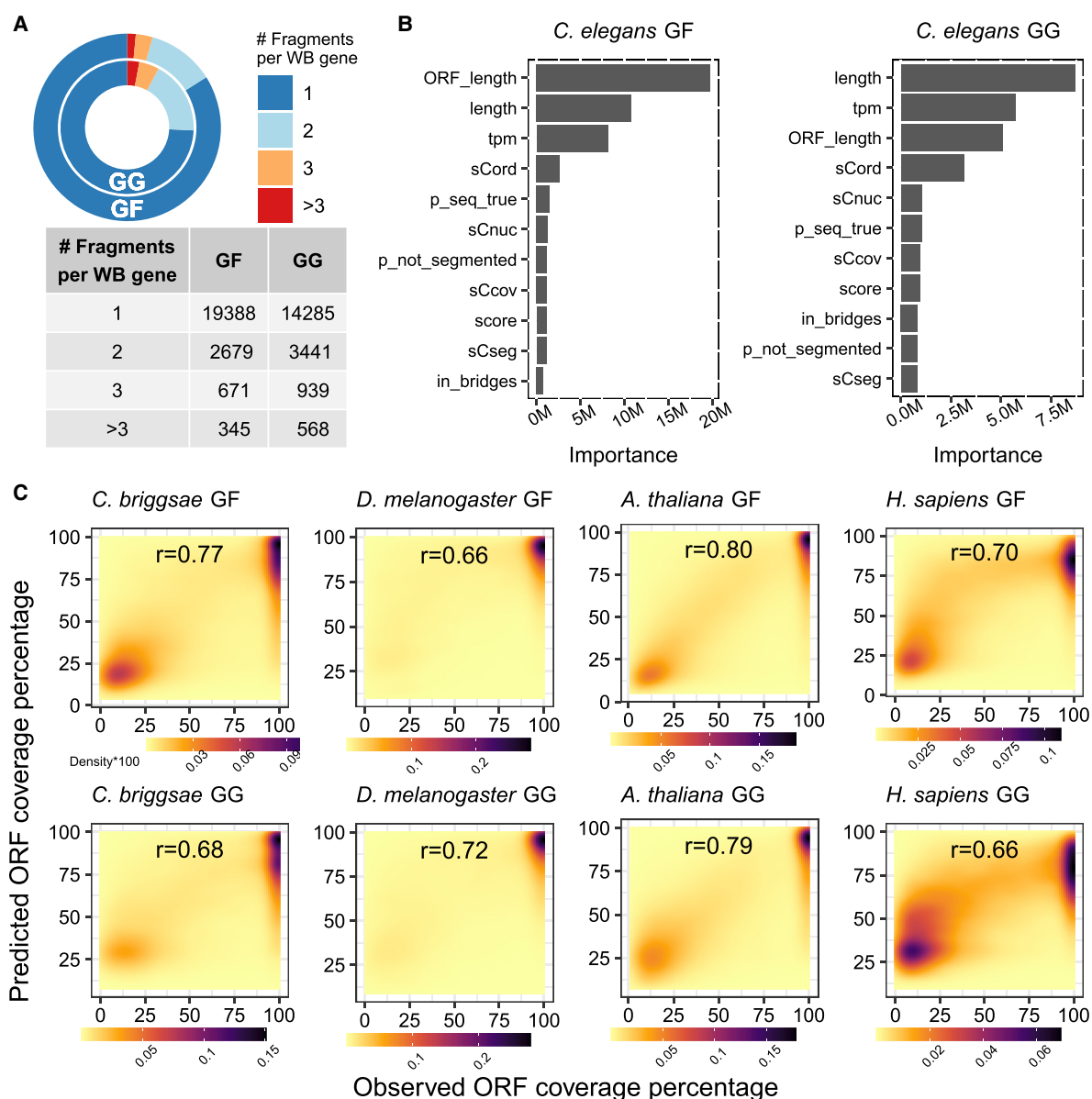


Figure 2. Benchmarking machine-learning-assisted filtering of fragmented transcripts in assemblies. (A) Pie chart and table representing the fragmentation status of the *C. elegans* GF and the GG transcriptome assembly. Shown are the proportions of WormBase genes overlapping with Trinity assembled transcripts. (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).

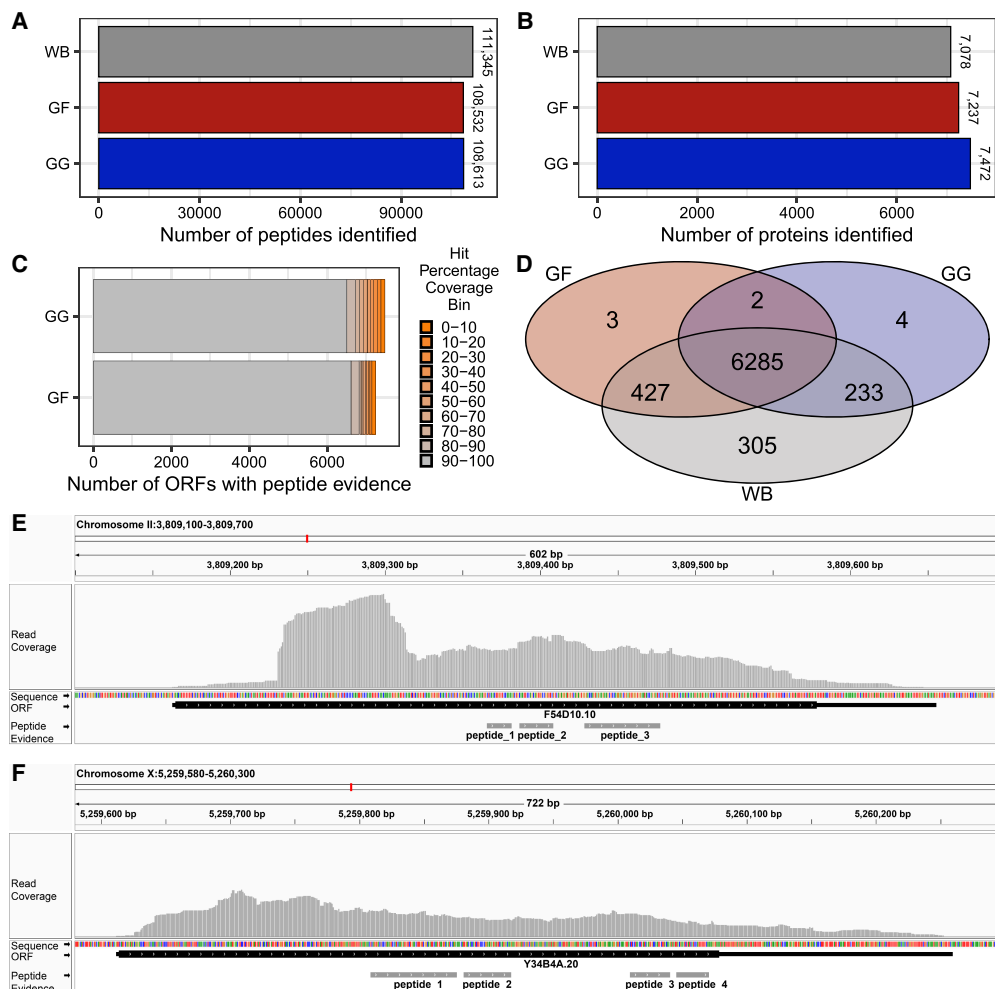


Figure 3. Proteotranscriptomics in *C. elegans*. (A) Number of peptides identified in the *C. elegans* WormBase, GF, and GG assemblies. (B) Number of individual proteins with peptide evidence for WormBase, GF, and GG proteomes. (C) Stacked bar plot of all ORFs of the GF and GG assembly with peptide evidence 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

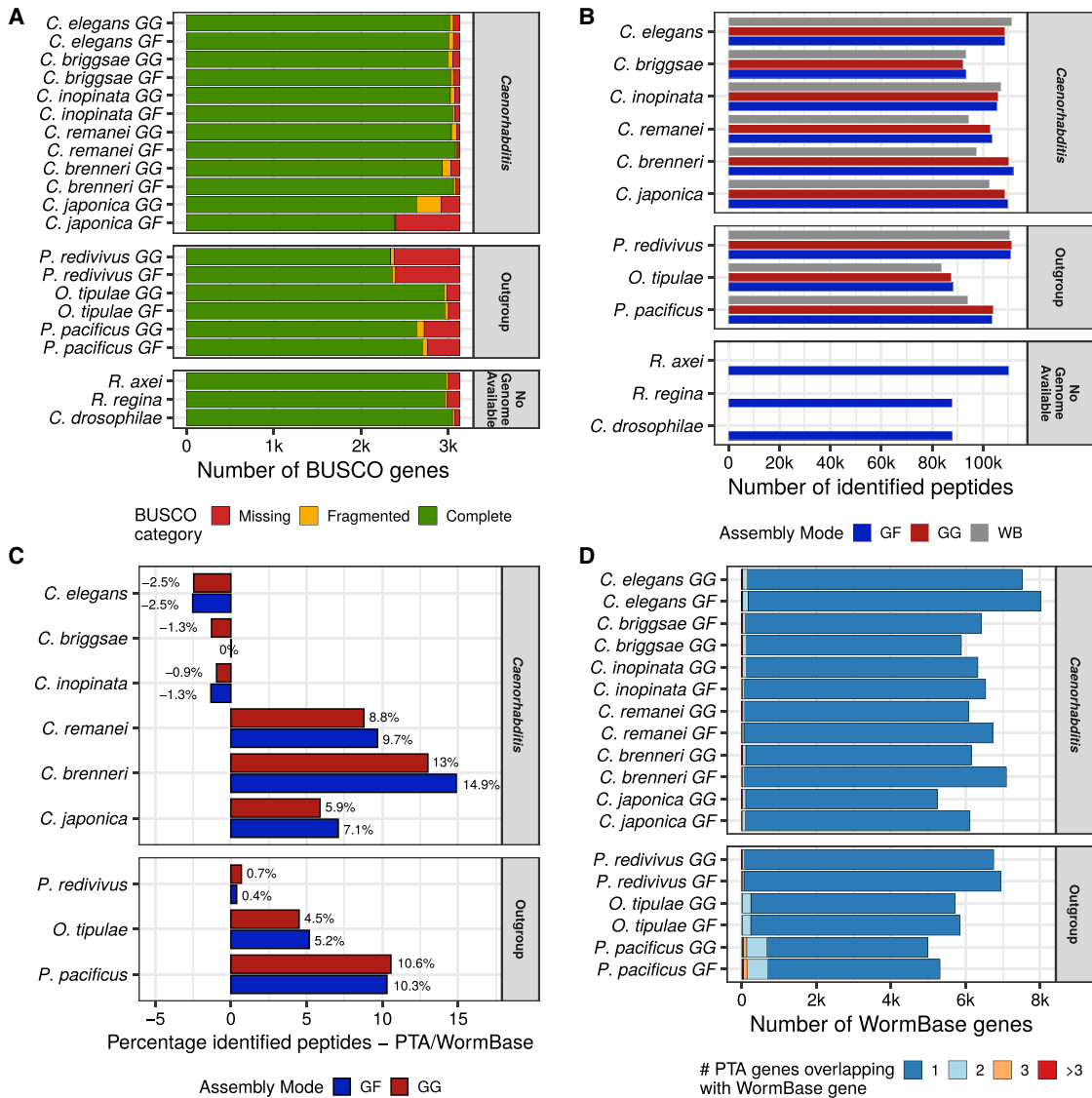


Figure 4. Proteotranscriptomics assembly (PTA) of 12 nematode species. (A) BUSCO metrics for all species and the two assembly modes, GF and GG. (B) Bar plot of mass spectrometry-identified peptides belonging to protein entries of the respective annotation source (GF, GG, and WB) for each species. (C) 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

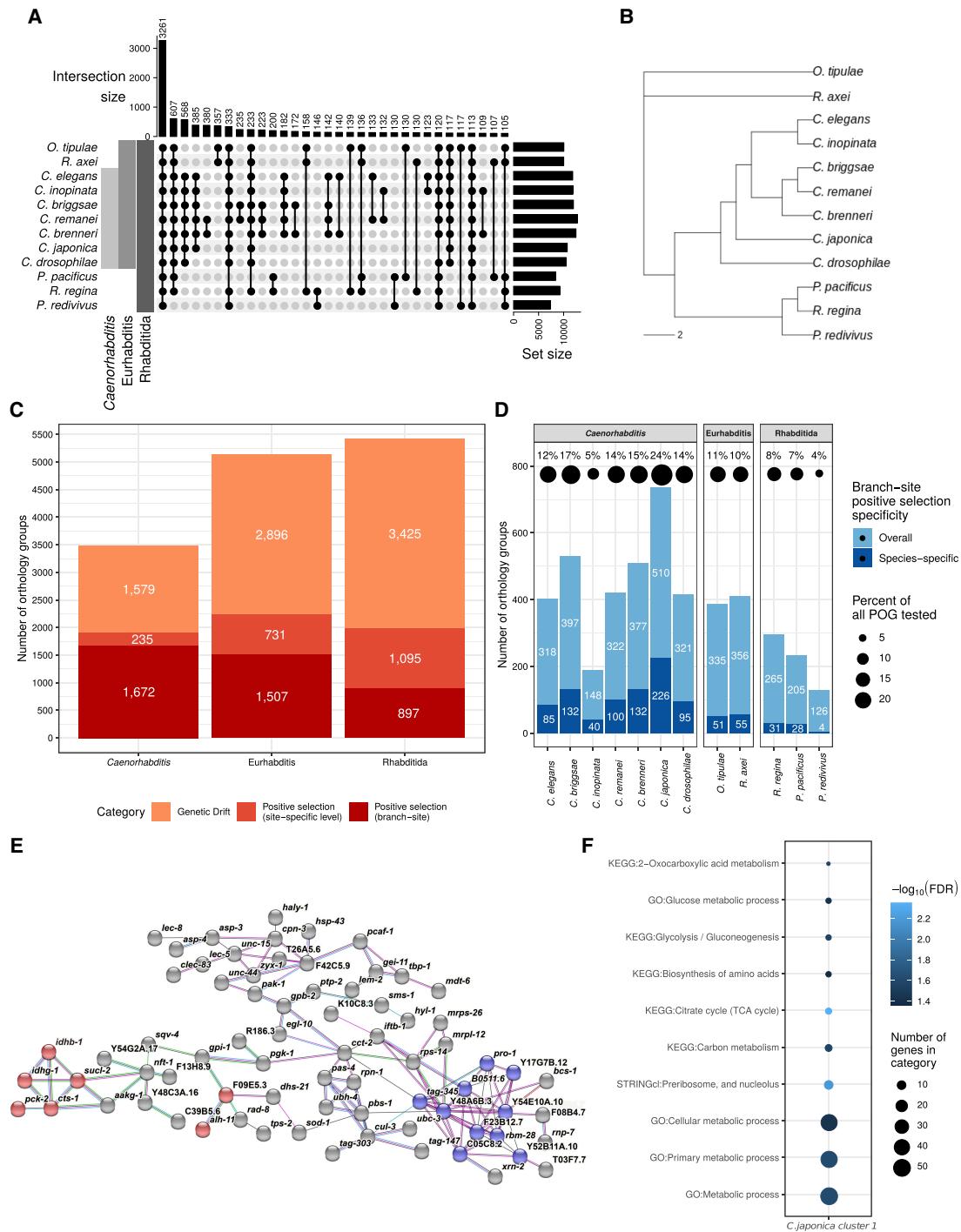


Figure 5. Orthology and phylogenetic relationships. (A) Upset plot depicting the number of orthology groups shared between different species. (B) Combined unrooted phylogenetic tree establishing the relationships between all studied species. The tree is based on individual gene trees of 1516 orthology 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 is 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 fragmenta-

tion 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 μ L of water in liquid nitrogen and stored at -80°C . For RNA isolation, 500 μ L 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 μ L 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 μ L water with liquid nitrogen and stored at -80°C . On the day of the protein isolation, samples were thawed and $2 \times$ lysis buffer (50 mM Tris-HCl, 300 mM NaCl, 3 mM MgCl_2 , 2 mM 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 $21,000g$ 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 $1 \times$ 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 iodoacetamide/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 500g, they were washed two times with 80% acetonitrile/0.1% formic acid and then with 0.1% formic acid for 5 min at 500g. The tryptic peptide samples were applied and centrifuged at 500g. 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 500g 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 ([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.](https://github.com/</p>
</div>
<div data-bbox=)

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 $2\Delta\text{LnL}$ difference of at least two between the two models for further analysis. We subsequently retrieved the *P*-value by comparing each $2\Delta\text{LnL}$ 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 ($\text{FDR} < 0.05$). 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 respec-

tive 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 $\text{FDR} < 0.1$ 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, Gevitzman 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**: bio209777. 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 defines 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 USA* **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/KNF121
- 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, Hatelye 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, Krijgsvelde 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/FCCELL.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, Garcia-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
- Romigüer J, Figueat 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, Dammernann 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, Bournsnel 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, Zucoli E, Zanin E, Mikeldadze-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, Govere 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, Westphal 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.



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. published online January 18, 2023

Access the most recent version at doi:[10.1101/gr.277070.122](https://doi.org/10.1101/gr.277070.122)

Supplemental Material <http://genome.cshlp.org/content/suppl/2023/01/18/gr.277070.122.DC1>

P<P Published online January 18, 2023 in advance of the print journal.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <https://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

Affordable, Accurate
Sequencing.



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

4.11.5 Article II: Supplementary information

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

Supplemental Data

Supplemental Figures S1-S12

- Supplemental Figure S1. Genome content and assembly contig N50 of 9 nematodes.
- Supplemental Figure S2. Confirmation levels of WormBase annotations.
- Supplemental Figure S3. Fragmented assembly visualization.
- Supplemental Figure S4. Expression profiles of two new ORFs in *C. elegans*.
- Supplemental Figure S5. BUSCO analysis in *P. pacificus* and *P. redivivus*.
- Supplemental Figure S6. Validation of *P. pacificus* fusion bias.
- Supplemental Figure S7. Overlap between WormBase, genome-guided and genome-free assemblies.
- Supplemental Figure S8. Network of *Caenorhabditis* specific genes.
- Supplemental Figure S9. Global levels of adaptive evolution.
- Supplemental Figure S10. Adaptive evolution in the *C. japonica* TCA cycle pathway.
- Supplemental Figure S11. Analyses of genome-guided dependent biases in *C. elegans*.
- Supplemental Figure S12. Correlation between transcript level and protein intensity and peptide sequence coverage.

Supplemental Tables S1-S10

- Supplemental Table S1. Summary information on strains, genome and gene features and transcriptome assemblies for all 12 species included in the study. (Separate files)
- Supplemental Table S2. Information on the number of assembled and evidenced ORFS and their overlap with WormBase. (Separate files)
- Supplemental Table S3. Fragmentation levels of genome-free (GF) and genome-guided (GG) transcriptome assemblies established by the comparison to WormBase annotations (version WB273) of the respective species. (Separate files)
- Supplemental Table S4. Detailed information on input features used for random forest training. (Separate files)
- Supplemental Table S5. Information on the assemblies of *C. briggsae*, *Drosophila melanogaster*, *Arabidopsis thaliana*, and human H1-hESC used for the machine learning benchmarking. (Separate files)
- Supplemental Table S6. Information on small proteins in *C. elegans* (GF+GG) with predicted completeness levels above 80% that are supported by at least 2 unique peptides. (Separate file).
- Supplemental Table S7. Information on presumably fused genes in *P. pacificus* in comparison to *C. elegans*. (Separate files)
- Supplemental Table S8. Enrichment analysis for ortholog groups encompassing all 12 species or all *Caenorhabditis* species. (Separate files)
- Supplemental Table S9. Table of orthology relations in the 12 species as produced by ProteinOrtho. (Separate files)
- Supplemental Table S10. Enrichment analysis results for species-specific lists of ORFs with signals of positive selection provided by STRINGdb. (Separate files)

Supplemental Material (.zip folder)

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

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

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

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

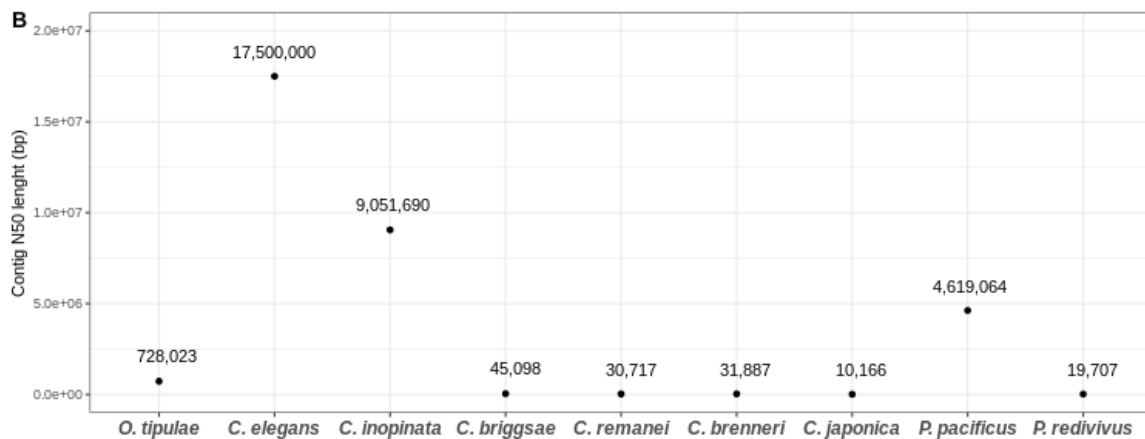
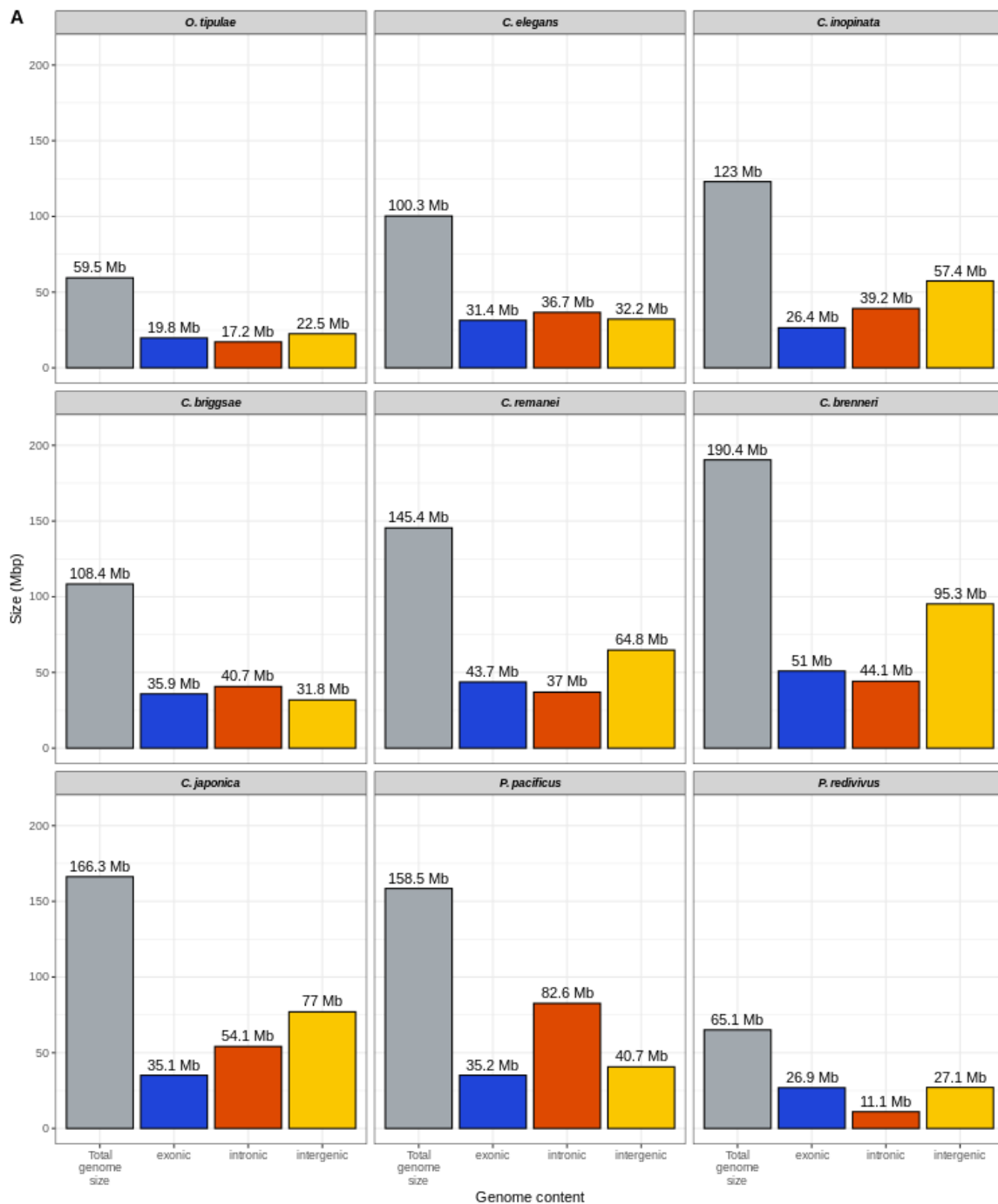
Transdecoder_ORF_predictions_pep_seq_evidenced.zip: protein sequence FASTA files of TransDecoder ORF predictions with mass spectrometry peptide evidence for all 12 species (separate files for genome-guided and genome-free transcriptome assembly).

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

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

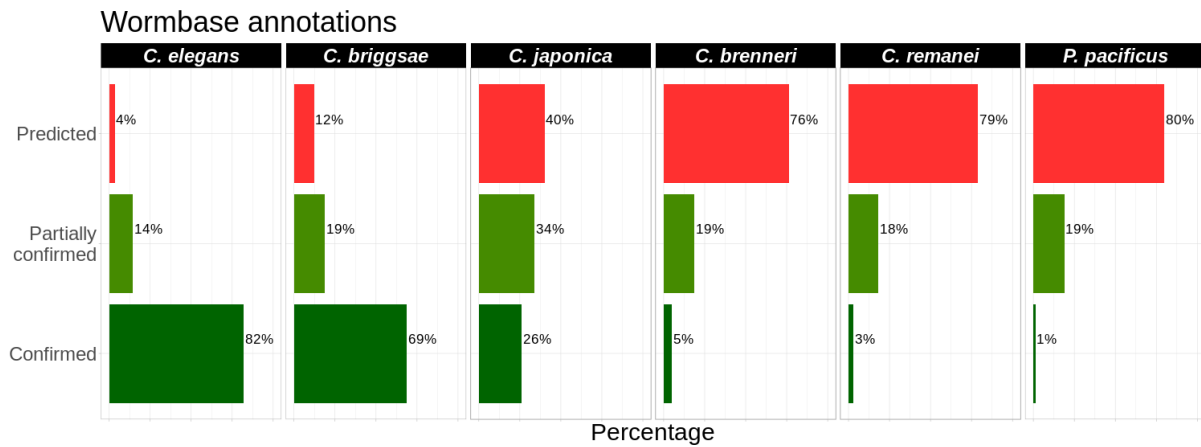
Blast_results_new_genes.pdf: BLASTP results for the two new *C. elegans* ORFs.

Supplemental Figure S1



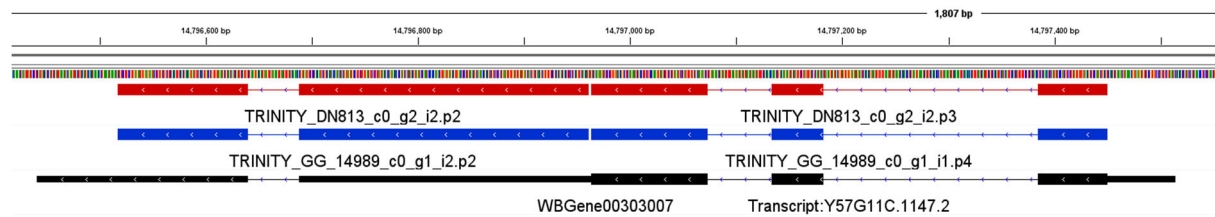
Supplemental Figure S1. Genome content and genome assembly contig N50 length of 9 nematodes. (A) Genome content extracted from WormBase genome assembly and gene annotation files (version WS273). Bar plots show the total genome size in gray and the proportions of exonic (blue), intronic (orange), and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

Supplemental Figure S2



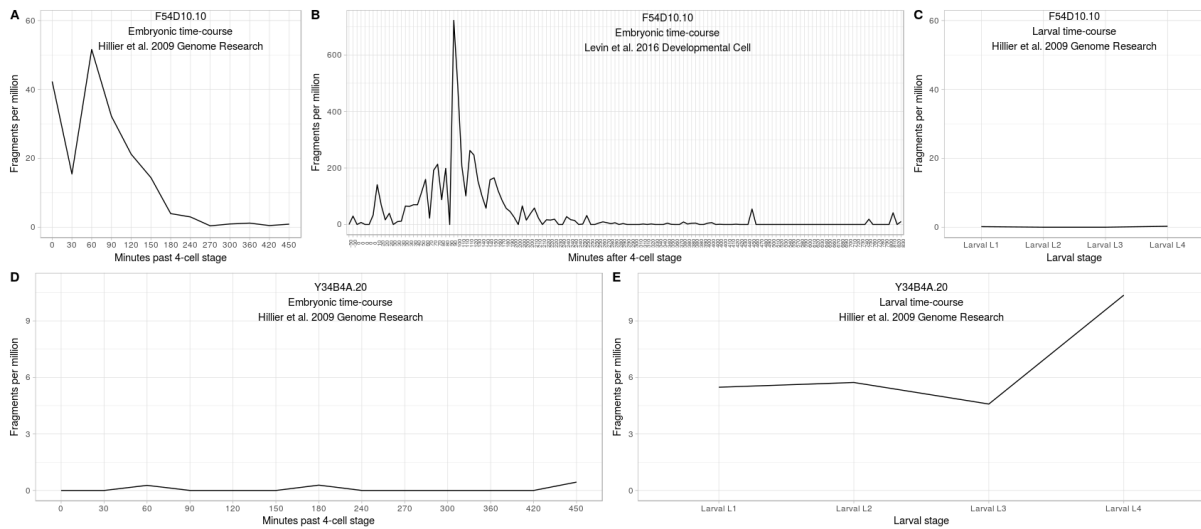
Supplemental Figure S2. Distribution of WormBase annotation confirmation level across all *Caenorhabditis* species. Categories are (1) predicted (red) - unsupported gene predictions, (2) partially confirmed (light green) - not all parts of the ORF are confirmed, and (3) confirmed (dark green) - all parts,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

Supplemental Figure S3



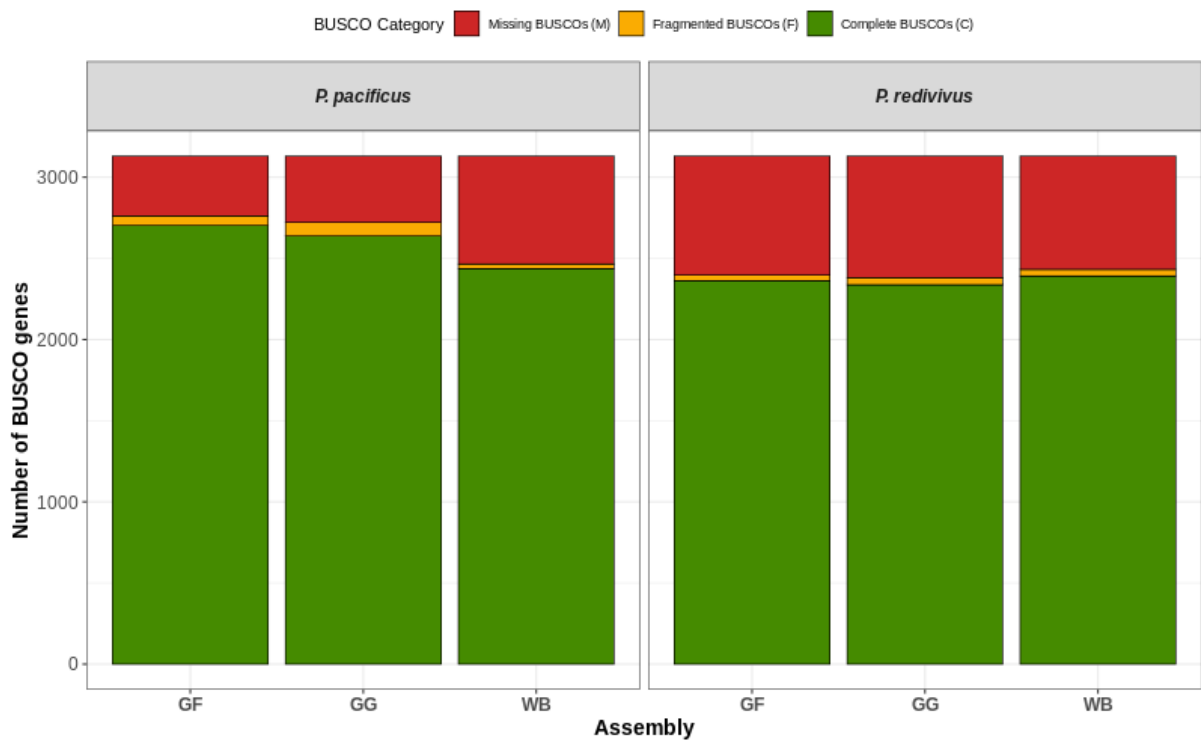
Supplemental Figure S3. Visualization of an example of a fragmented gene model via Integrative Genomics Viewer (IGV) browser. *C. elegans* GF (red) and GG (blue) assembled transcripts were mapped to the *C. elegans* genomic sequence and are shown side by side with the respective *C. elegans* WormBase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

Supplemental Figure S4



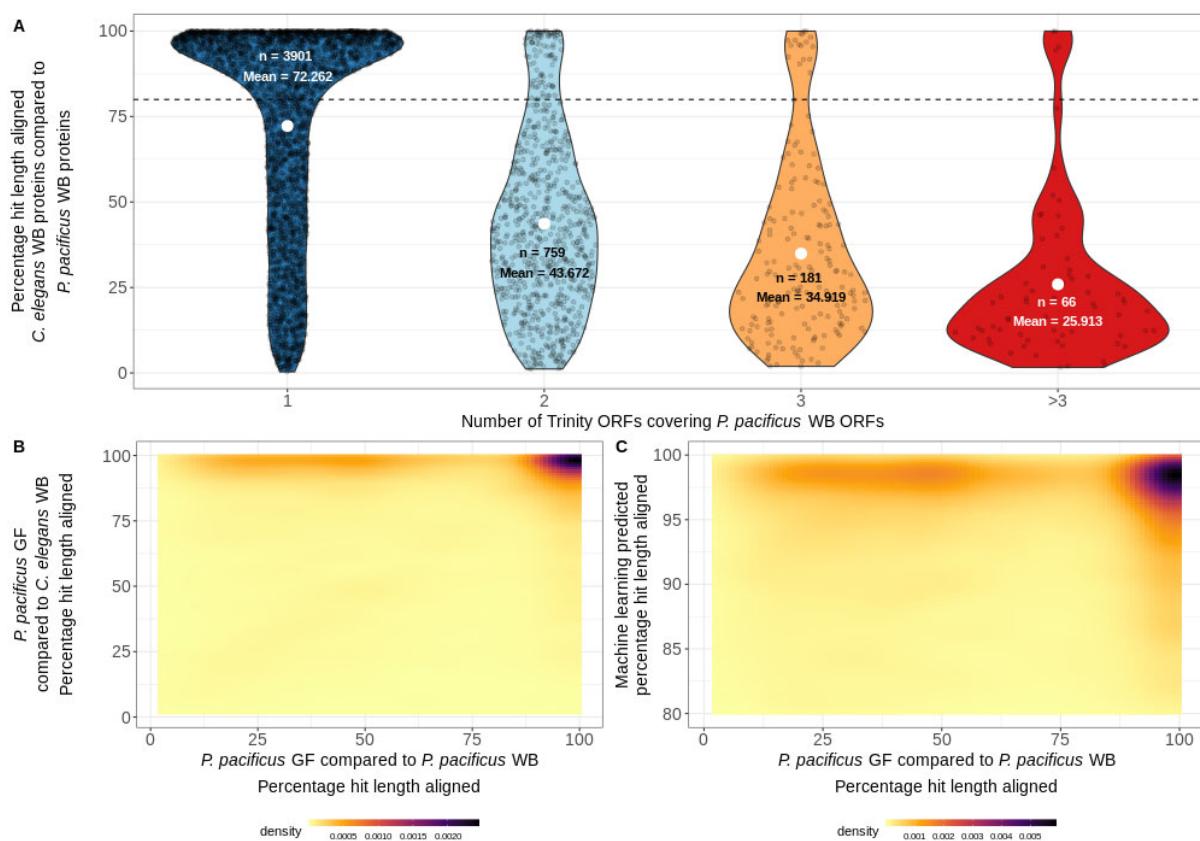
Supplemental Figure S4. Expression profiles of F54D10.10 and Y34B4A.20 during developmental stages of *C. elegans*. (A) F54D10.10 transcript shows expression during embryonic developmental time-course. (B) Validation of F54D10.10 embryonic expression in an additional embryonic transcriptome time-course.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

Supplemental Figure S5



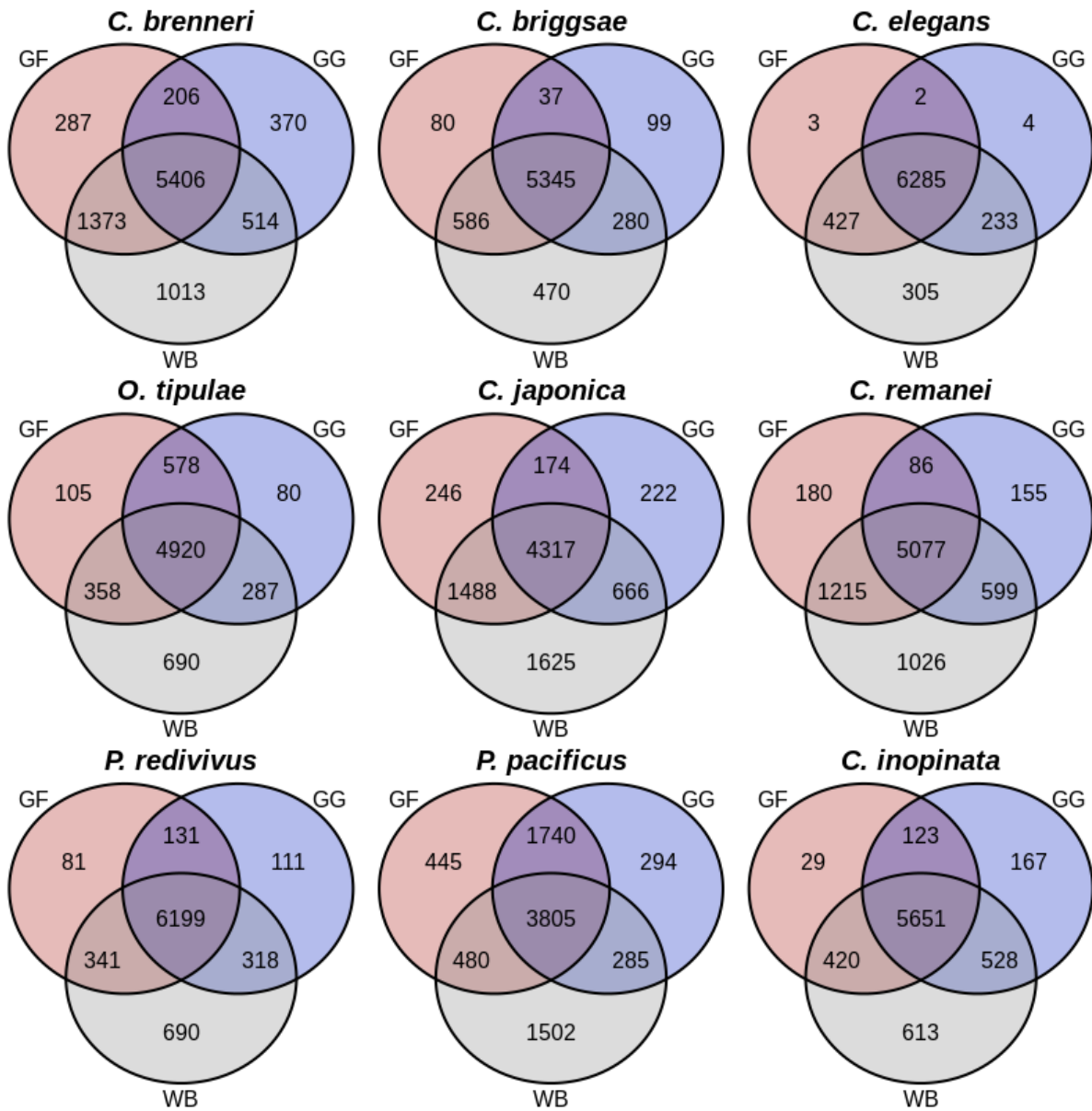
Supplemental Figure S5. Results of BUSCO analysis comparing GF and GG assemblies with the current WormBase annotation of *P. pacificus* and *P. redivivus*. The y-axis represents the counted number of BUSCO genes and the x-axis shows different evaluated assemblies. Green: complete and single-copy genes;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

Supplemental Figure S6



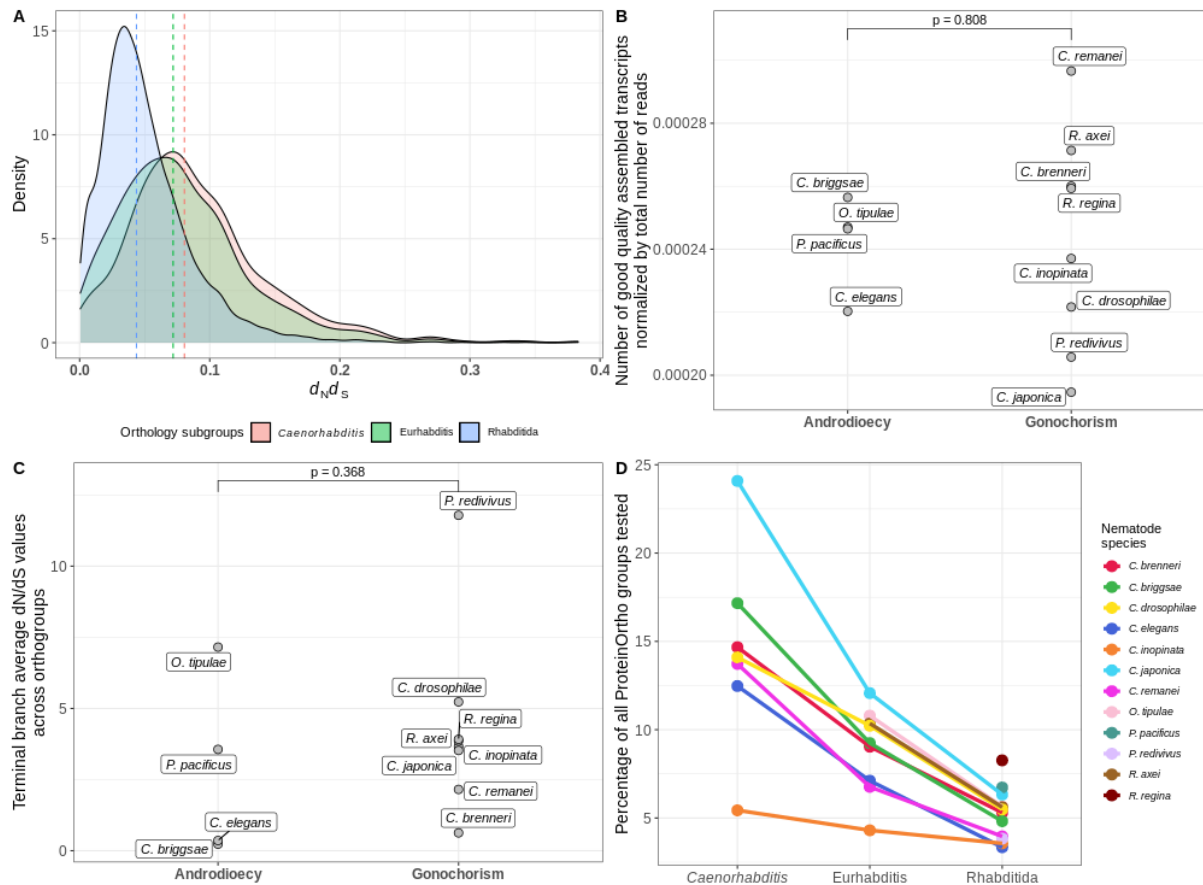
Supplemental Figure S6. Validation of *P. pacificus* fusion bias. (A) Percentage hit length of WormBase *P. pacificus* proteins when compared to WormBase *C. elegans* orthologs established by blastp. *P. pacificus* proteins were grouped by the number of proteins needed to cover the same protein sequence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

Supplemental Figure S7



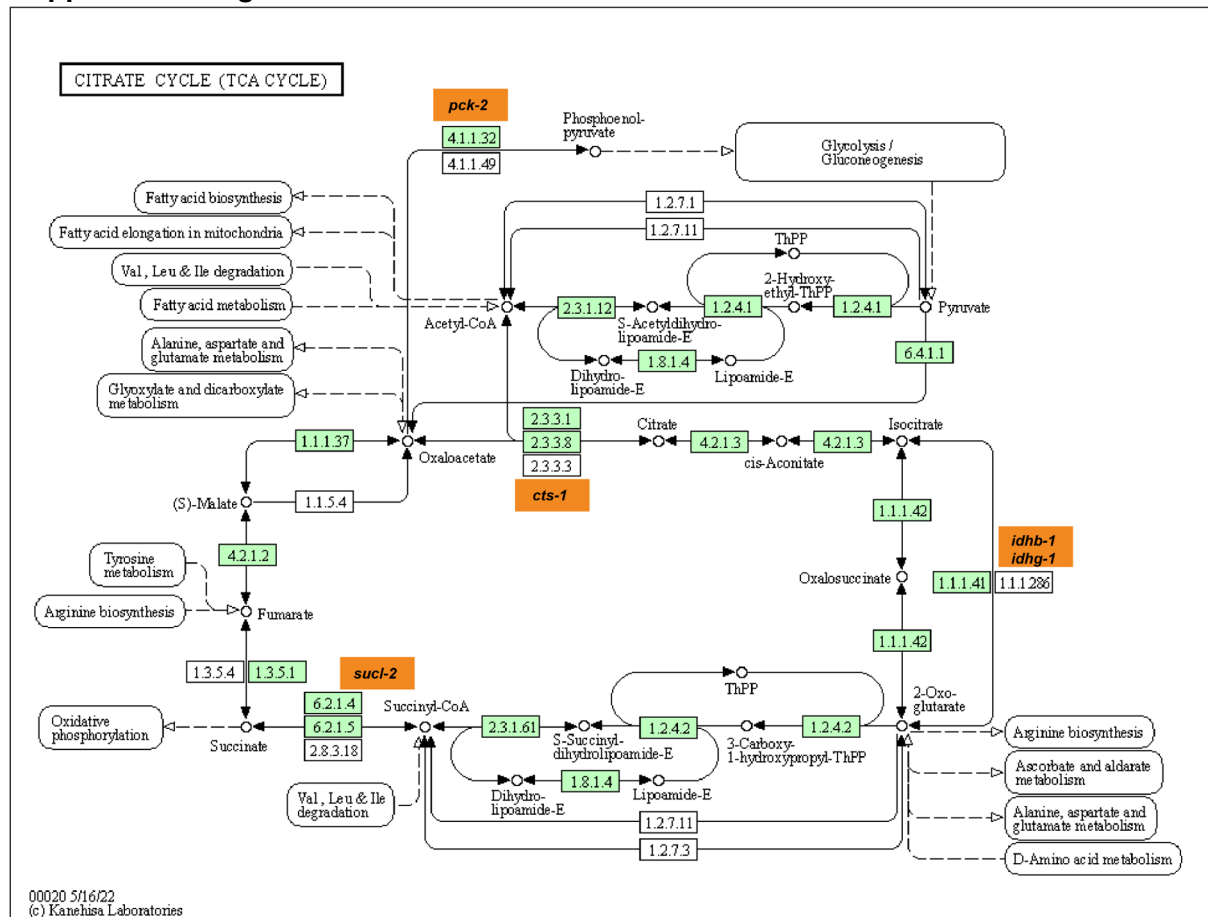
Supplemental Figure S7. Venn diagrams depicting the overlap between the identified proteins of WB (WormBase in gray), GF (genome-free in red), and GG (genome-guided in blue) proteomes for each studied species.

Supplemental Figure S9



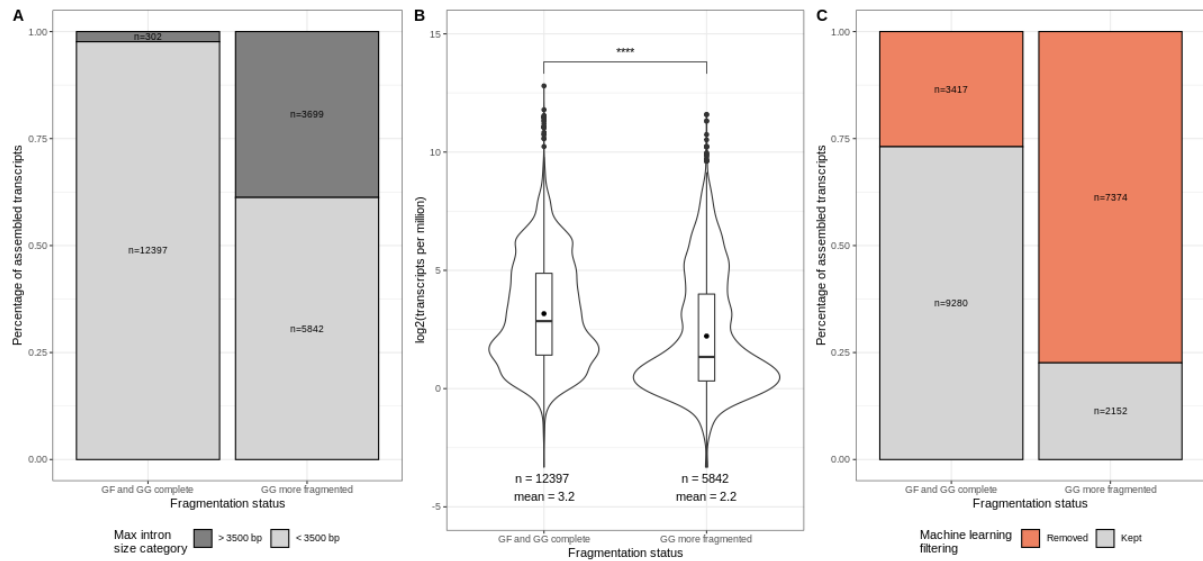
Supplemental Figure S9. Different levels of adaptive evolution detected in a phylogeny of nematodes. (A) Density plot of M0 model d_N/d_S (ω) values calculated for 5,417 orthologs in 12 nematodes species, evaluated for Rhabditida (blue), Eurhabditis (green), and *Caenorhabditis* (red). The median of each group is represented with a dashed line. All distributions show high levels of purifying selection ($\omega < 0.1$) in the majority of the codon sites. The differences in the medians and shift in the distributions of the values between the different groups emphasize the decrease in the detection sensitivity of adaptive evolution with an increasing degree of divergence between species (*Caenorhabditis* > Eurhabditis > Rhabditida). (B) Assembly efficiency measured as the number of assembled transcripts that pass the machine learning completeness prediction of 80% normalized by the total number of sequenced reads used for the assembly is shown for species divided into gonochoristic and androdioecious mode of reproduction. Due to missing genome annotations and uncertainty regarding the quality of some of the existing assemblies only genome-free assembled transcripts are represented. (C) Terminal branch average d_N/d_S values across 1-to-1 orthogroups are shown for species divided into gonochoristic and androdioecious mode of reproduction. (D) Percentages of orthologous groups under positive selection, grouped by subsets of species included in the analysis - Rhabditida, Eurhabditis, and *Caenorhabditis*.

Supplemental Figure S10



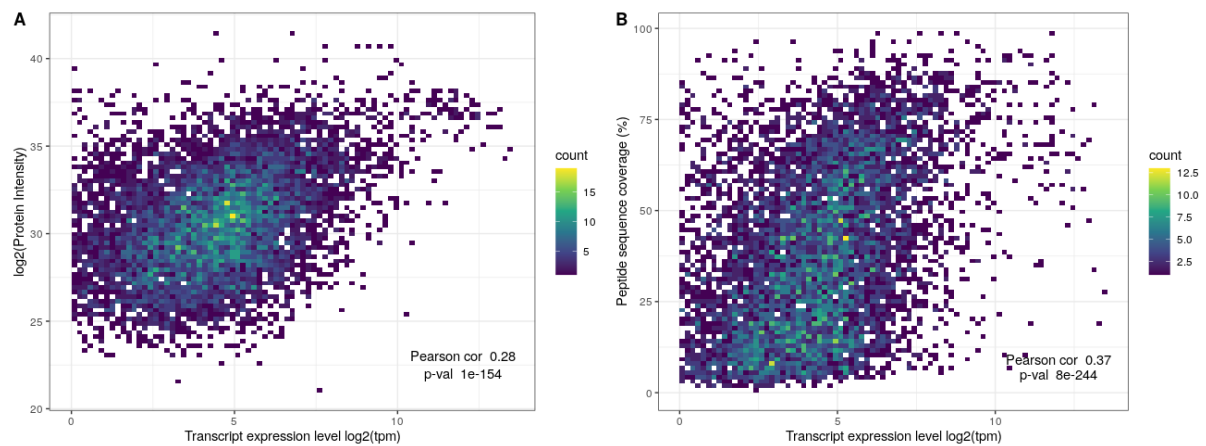
Supplemental Figure S10. Tricarboxylic acid cycle (TCA) KEGG pathway. TCA cycle genes under positive selection in *C. japonica* are highlighted in orange. Pathway diagram was adapted from <https://www.kegg.jp/pathway/ce100020>.

Supplemental Figure S11



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

Supplemental Figure S12



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

4.12 Article III: AlexandrusPS - a user-friendly pipeline for genome-wide positive selection analysis

4.12.1 Summary

AlexandrusPS is a user-friendly pipeline designed for genome-wide positive selection analysis, implemented as a combination of Perl, R, and shell scripts running in a Linux/UNIX environment. The pipeline, provided as an open-source solution, is packed in a Docker image to minimize the need for local installation and requires only CDS and peptide FASTA files as input. With its automated process, **AlexandrusPS** generates orthology relationships, sequence alignments and phylogenetic trees. It then performs site-specific (SSM), branch (BM) and branch-site (BSM) positive selection analyses, and produces four main output files including orthology relationships, positive selection results, and all intermediate files (sequence alignment, phylogenetic tree). **AlexandrusPS** offers significant advantages over other programs with its user-friendly interface, efficient execution of CodeML models (SSM, BM, and BSM), and the ability to run on desktop computers in a parallel manner without the need for high-performance computing systems. A detailed manual can be found on the **AlexandrusPS** GitHub page (<https://github.com/alejocn5/AlexandrusPS>).

4.12.2 Zusammenfassung

AlexandrusPS ist eine benutzerfreundliche Pipeline, die für die genomweite positive Selektionsanalyse entwickelt wurde. Sie ist als Kombination aus Perl-, R- und Shell-Skripten implementiert und wird in einer Linux/UNIX-Umgebung ausgeführt. Die als Open-Source-Lösung bereitgestellte Pipeline ist in ein Docker-Image gepackt, um die Notwendigkeit einer lokalen Installation zu minimieren, und erfordert nur die CDS und Peptid FASTA Sequenz Dateien als Eingabe. Mit seinem automatisierten Prozess generiert **AlexandrusPS** Orthologiebeziehungen, Sequenzalignments und phylogenetische Bäume. Es führt dann site-spezifische (SSM), branch-spezifische (BM) und branch-site-spezifische (BSM) positive Selektionsanalysen durch und erstellt vier Ausgabedateien, einschließlich orthologischer Beziehungen, positiver Selektionsergebnisse und aller Zwischendateien (Sequenzalignment, phylogenetischer Baum). **AlexandrusPS** ermöglicht mit seiner benutzerfreundlichen Oberfläche die effiziente parallele Auswertung von **CodeML**-Modellen (SSM, BM und BSM) auf Desktop-Computern ohne Bedarf an Hochleistungs-Computersystemen. Eine ausführliche Anleitung befindet sich auf der GitHub-Seite **AlexandrusPS** (<https://github.com/alejocn5/AlexandrusPS>)

4.12.3 Statement of Contribution

4.12.4 Article III: Main Text

On Preparation - Not Submitted

AlexandrusPS - a user-friendly pipeline for genome-wide positive selection analysis

Alejandro Ceron-Noriega¹, Vivien A. C. Schoonenberg¹, Falk Butter^{1,2*} and Michal Levin^{1,2*}

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

²These authors contributed equally.

*Correspondence: m.levin@imb.de (ML), f.butter@imb.de (FB).

Abstract

Motivation: AlexandrusPS is a high-throughput pipeline that overcomes technical challenges when conducting genome-wide positive selection analyses on large sets of nucleotide and protein sequences. These challenges include i) the execution of an accurate orthology prediction as a precondition for positive selection analysis, ii) preparing and organizing configuration files for CodeML and iii) generating an output that is easy to interpret including all maximum likelihood and log likelihood test results. The only input needed from the user are the CDS and peptide FASTA files of all proteins of interest. Provided in a Docker image no program or module installation is required. The pipeline runs on a desktop computer making it easily applicable.

Availability: AlexandrusPS is available via GitHub (<https://github.com/alejocn5/AlexandrusPS>) and as an easy-to-use Docker container

Supplementary information: Supplementary data are available at <https://github.com/alejocn5/AlexandrusPS> online.

1 Introduction

1 Introduction The evolution of protein sequences is manifested by constraining changes (purifying selection) or by favoring the fixation of alleles that confer fitness advantage (positive selection) (Maldonado *et al.* (2016)). An essential metric to detect the selection type driving the evolution of protein-coding sequences is the nucleic acid and amino acid substitution rate, namely the nonsynonymous (d_N) to synonymous (d_S) substitution rate ratio ($\omega = d_N/d_S$). This measure has proven to be useful for understanding different evolutionary processes in comparative genomics (bookfelsenstein2004inferring, (Huelsenbeck and Rannala (1997)), (Sánchez *et al.* (2011)), (Parker *et al.* (2013)), (Li *et al.* (2014)), (Glover *et al.* (2019)), (Pan *et al.* (2013)), (Chuang and Li (2004)) (Stark *et al.* (2007)), (Policarpo *et al.* (2021)), (Liu *et al.* (2019)), (Bast *et al.* (2018)) (Clark *et al.* (2003)) (Fedorova *et al.* (2008)) (Egan *et al.* (2008)), (Forni *et al.* (2021)). Such evolutionary analyses have profited from massive amounts of data derived from Next Generation Sequencing (NGS) technologies, making comparative genomics analyses more attainable. The enormous quantity of such data provides a valuable resource for researchers, but as the number of genomes continues to grow, downstream

analyses have become increasingly challenging. This problem has led to the need to develop specialized, efficient and user-friendly bioinformatics tools that can help researchers in downstream tasks (Koepfli *et al.* (2015)).

One of the most popular bioinformatics tools applying maximum likelihood (ML) based models in evolution research to test the ratio between nonsynonymous and synonymous substitutions ($\omega = d_N/d_S$) for multiple orthologous protein-coding sequences is CodeML (Yang (2007)). CodeML is implemented in the PAML (Phylogenetic Analysis by Maximum Likelihood) program package (Yang (2007)), (Maldonado *et al.* (2016)). While the program is statistically robust and highly accurate in examining selective pressure (Maldonado *et al.* (2016)), (Zhai *et al.* (2012)), (Gharib and Robinson-Rechavi (2013)), (Macías *et al.* (2020)) CodeML also faces limitations: i) Being executed on a single CPU renders operations on large sets of sequences highly time-consuming, driving the need for accessibility to high-performance computers. ii) Each individual execution needs to be manually performed by the user. iii) The execution requires a preceding accurate orthology analysis, which itself is challenging and can introduce errors to the analysis if not performed properly and iv) CodeML provides output that is difficult to interpret especially for inexperienced users (Steffen *et al.* (2022)), (Maldonado

et al. (2013)), (Maldonado et al. (2016)).

To support less experienced users and minimize the manual operation of CodeML several programs have emerged: JCoDA (Steinway et al. (2010)), Armadillo (Lord et al. (2012)), PAMLX (Xu and Yang (2013)), IMPACT-S (Maldonado et al. (2014)), PSP (Su et al. (2013)), PhyleasProg (Busset et al. (2011)), and Selecton (Stern et al. (2007)). These programs use graphical interfaces or web-server implementations for single-gene family analysis, however, they are not suitable for streamlined operation of CodeML for multiple analyses. Some additional software attempts to solve these large-scale analysis challenges: VESPA (Webb et al. (2017)), IDEA (Egan et al. (2008)), and POTION (Hongo et al. (2015)), but these programs still have certain shortcomings: i) The installation is complex. ii) They depend on large computational infrastructure such as high-performance computers (HPC) and iii) they require advanced programming skills of the user.

Here we introduce *AlexandrusPS*, a high-throughput user-friendly pipeline designed to simplify the automated operation of established CodeML protocols for researchers with less bioinformatics experience. Containerized in a Docker image, *AlexandrusPS* was developed as a single command pipeline minimizing user intervention, in both installation as well as execution. The pipeline provides a well-organized output table including all relevant results for drawing conclusions. All intermediate data such as the results of the orthology analysis as well as multiple sequence alignments are also retained. To enable full analysis flexibility for more experienced researchers, *AlexandrusPS* is an open source software and thus enables modifications of parameters in all major configuration files.

2 Implementation

2.1 AlexandrusPS functionality

AlexandrusPS is a pipeline consisting of Perl and R scripts called by a main shell script and is implemented in a Docker image (Docker (2020)). The only input needed from the user are FASTA files of all CDS and amino acid sequences of all target proteins. *AlexandrusPS* will then predict orthologous gene clusters (OGC) that are used for the analysis of molecular adaptive evolution with various CodeML models. These results are then used for likelihood ratio tests (LRTs) to determine whether the models reflect diversifying selection. For this, the log-likelihood score $2\Delta\text{LnL}$ between any two models is calculated. Subsequently, the P-value is determined by comparing each $2\Delta\text{LnL}$ against the χ^2 distribution using the respective degrees-of-freedom (DoF) for each model pair. Significant LRT results (FDR < 0.05) indicate a significant difference between the two models and thus imply an evolutionary explanation for these differences.

The main workflow (Fig. 1) is composed of four steps: i) Orthology prediction by ProteinOrtho (Lechner et al. (2011)); ii) Multiple amino acids and codon alignment by PRANK (Suyama et al. (2006)) and pal2nal (Suyama et al. (2006)); iii) site-specific model calculations by CodeML (Yang (2007)); iv) branch and branch-site-specific model calculations by CodeML (Yang (2007)).

2.2 AlexandrusPS input files

2.2.1 FASTA files of all proteins of interest

For each species included in the analysis two FASTA files are needed: one with the sequences of amino acids and the other with the respective CDS sequences. Both files should contain the same number of sequences and their headers must be identical. *AlexandrusPS* can perform genome-wide analysis, i.e. it can analyze all protein groups from whole genome gene predictions/transcriptomes across multiple species. An example

data set is provided with the pipeline that enables testing of the proper functionality of the pipeline.

2.3 AlexandrusPS output files

Site-Specific Models (SSM): The CodeML output files are parsed into a CSV file. This file contains all OGCs organized in rows. Columns include OGC_ID, Species included in the OGC and ML results for all models with the respective metrics such as likelihood (lnL), the number of parameters (np), ω (d_N/d_S), degrees of freedom (DoF), log likelihood value (lnL), likelihood ratio tests (LRT) and positively selected sites (PSS).

Branch models and Branch-site (BM and BSM): Results of the LTR-based branch/branch-site model analyses (null model (H0) and alternative model (H1) of the branch-site test for the OGC with significant signals of site-specific diversifying selection are written into final easily interpretable results files.

2.4 AlexandrusPS execution and paralleling

CodeML is limited to running one analysis per CPU. One of the benefits of *AlexandrusPS* is that the user can run multiple CodeML analyses on distinct CPUs in parallel. The number of CPUs used by *AlexandrusPS* can be adjusted by the user.

2.5 AlexandrusPS proof-of-principle

AlexandrusPS was successfully applied to perform a large-scale positive selection analysis using proteotranscriptomics data across 12 nematode species including up to 5,400 orthology groups (Ceron-Noriega et al. (2023)). This extensive phylogenetic systems analysis included 77,000 protein sequences, was executed on a tabletop PC of 16 CPUs and finished within 7 days. The analysis allowed interesting new insights into evolutionary processes of this metazoan group.

3 Conclusion

AlexandrusPS is a PC-based pipeline that is packed in a Docker image to avoid the need for local installation of any modules or programs. It is provided as an open-source pipeline that allows the use of various CodeML models for molecular adaptive evolution (SSM, BM, and BSM) in parallel. It can run with default parameters, as it is based on standard protocols that allow an analysis of large-scale, genome-wide datasets. Users are only required to provide the CDS and peptide FASTA files of the proteins of interest. With its usage simplicity, *AlexandrusPS* offers distinct advantages over other programs. *AlexandrusPS* automatically generates orthology relationships and identifies optimal orthology groups for positive selection analysis to avoid problems such as paralog introduction. It also generates a gene tree of each OGC and organizes, executes and extracts all pertinent information from CodeML outputs. This completely automates the analysis with no need for intervention by the user. *AlexandrusPS* generates four main outputs: Orthology relationships, site-specific positive selection results, branch and branch-site positive selection results, along with all intermediate files for each OGC. These intermediate files enable manual repetition of certain analyses for any individual OGC without having to repeat the entire process. *AlexandrusPS* allows users to run CodeML protocols on a desktop computer in an automated parallel manner, facilitating high-throughput analyses without the need for high-performance computer systems. We successfully applied *AlexandrusPS* in the genome-wide investigation of positive selection in a phylogeny of 12 nematode species and received highly interesting results (Ceron-Noriega et al. (2023)). We believe that this implementation will empower many more researchers to explore positive selection in any species compendium of interest.

4 Funding

This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK2526/1 – Projectnr. 407023052. Conflict of Interest: none declared.

References

- Bast, J. *et al.* (2018). Consequences of asexuality in natural populations: insights from stick insects. *Molecular biology and evolution*, **35**(7), 1668–1677.
- Busset, J. *et al.* (2011). Phyleasprog: a user-oriented web server for wide evolutionary analyses. *Nucleic acids research*, **39**(suppl_2), W479–W485.
- Ceron-Noriega, A. *et al.* (2023). Nematode gene annotation by machine-learning-assisted proteotranscriptomics enables proteome-wide evolutionary analysis. *Genome Research*.
- Chuang, J. H. and Li, H. (2004). Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS biology*, **2**(2), e29.
- Clark, A. G. *et al.* (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**(5652), 1960–1963.
- Docker, I. (2020). Docker. *linea*[Junio de 2017]. Disponible en: <https://www.docker.com/what-docker>.
- Egan, A. *et al.* (2008). Idea: interactive display for evolutionary analyses. *BMC bioinformatics*, **9**(1), 1–9.
- Fedorova, N. D. *et al.* (2008). Genomic islands in the pathogenic filamentous fungus *aspergillus fumigatus*. *PLoS genetics*, **4**(4), e1000046.
- Forni, G. *et al.* (2021). Base: A novel workflow to integrate nonubiquitous genes in comparative genomics analyses for selection. *Ecology and Evolution*, **11**(19), 13029–13035.
- Gharib, W. H. and Robinson-Rechavi, M. (2013). The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in gc. *Molecular biology and evolution*, **30**(7), 1675–1686.
- Glover, N. *et al.* (2019). Advances and applications in the quest for orthologs. *Molecular biology and evolution*, **36**(10), 2157–2164.
- Hongo, J. A. *et al.* (2015). Potion: an end-to-end pipeline for positive darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. *BMC genomics*, **16**(1), 1–16.
- Huelsenbeck, J. P. and Rannala, B. (1997). Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science*, **276**(5310), 227–232.
- Koepfli, K.-P. *et al.* (2015). The genome 10k project: a way forward. *Annu. Rev. Anim. Biosci.*, **3**(1), 57–111.
- Lechner, M. *et al.* (2011). Proteinortho: detection of (co-) orthologs in large-scale analysis. *BMC bioinformatics*, **12**(1), 1–9.
- Li, C. *et al.* (2014). Two antarctic penguin genomes reveal insights into their evolutionary history and molecular changes related to the antarctic environment. *GigaScience*, **3**(1), 2047–217X.
- Liu, A. *et al.* (2019). Convergent degeneration of olfactory receptor gene repertoires in marine mammals. *BMC genomics*, **20**(1), 1–14.
- Lord, E. *et al.* (2012). Armadillo 1.1: an original workflow platform for designing and conducting phylogenetic analysis and simulations. *PLoS one*, **7**(1), e29903.
- Macías, L. G. *et al.* (2020). Gwidecodeml: a python package for testing evolutionary hypotheses at the genome-wide level. *G3: Genes, Genomes, Genetics*, **10**(12), 4369–4372.
- Maldonado, E. *et al.* (2013). Easer: Ensembl easy sequence retriever. *Evolutionary Bioinformatics*, **9**, EBO–S11335.
- Maldonado, E. *et al.* (2014). Impact_s: integrated multiprogram platform to analyze and combine tests of selection. *PLoS one*, **9**(10), e96243.
- Maldonado, E. *et al.* (2016). Lmap: lightweight multigene analyses in paml. *BMC bioinformatics*, **17**(1), 1–11.
- Pan, D. *et al.* (2013). Genome-wide detection of selective signature in chinese holstein. *PLoS one*, **8**(3), e60440.
- Parker, J. *et al.* (2013). Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*, **502**(7470), 228–231.
- Policarpo, M. *et al.* (2021). Contrasting gene decay in subterranean vertebrates: insights from cavefishes and fossorial mammals. *Molecular biology and evolution*, **38**(2), 589–605.
- Sánchez, R. *et al.* (2011). Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic acids research*, **39**(suppl_2), W470–W474.
- Stark, A. *et al.* (2007). Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature*, **450**(7167), 219–232.
- Steffen, R. *et al.* (2022). papaml: An improved computational tool to explore selection pressure on protein-coding sequences. *Genes*, **13**(6), 1090.
- Steinway, S. N. *et al.* (2010). Jcoda: a tool for detecting evolutionary selection. *BMC bioinformatics*, **11**(1), 1–9.
- Stern, A. *et al.* (2007). Selecton 2007: advanced models for detecting positive and purifying selection using a bayesian inference approach. *Nucleic acids research*, **35**(suppl_2), W506–W511.
- Su, F. *et al.* (2013). Psp: rapid identification of orthologous coding genes under positive selection across multiple closely related prokaryotic genomes. *BMC genomics*, **14**(1), 1–10.
- Suyama, M. *et al.* (2006). Pal2nal: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research*, **34**(suppl_2), W609–W612.
- Webb, A. E. *et al.* (2017). Vespa: very large-scale evolutionary and selective pressure analyses. *PeerJ Computer Science*, **3**, e118.
- Xu, B. and Yang, Z. (2013). Pamlx: a graphical user interface for paml. *Molecular biology and evolution*, **30**(12), 2723–2724.
- Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, **24**(8), 1586–1591.
- Zhai, W. *et al.* (2012). Looking for darwin in genomic sequences—validity and success of statistical methods. *Molecular biology and evolution*, **29**(10), 2889–2893.

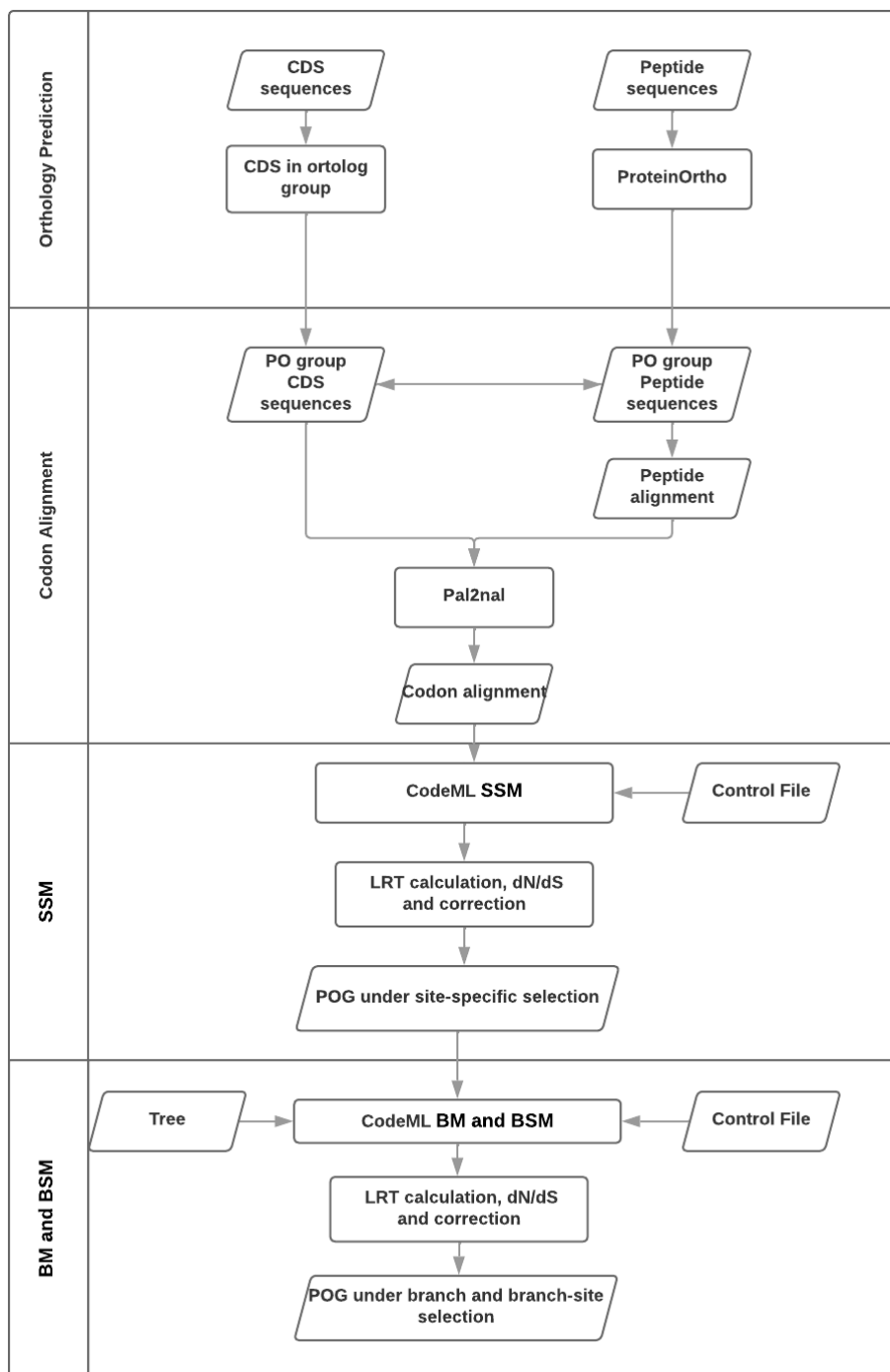


Figure 1: AlexandrusPS workflow. Flowchart describing AlexandrusPS workflow which sequentially combines four steps to finally execute CodeML and collect results. PO = ProteiOrtho; SSM = Specific Site Model; BM = Branch Model; BSM = Branch Site Model; LRT = Likelihood Ratio Test; POG = ProteinOrtho Group.

4.12.5 Article III: Supplementary information

GitHub: AlexandrusPS

Contents

1	Introduction	2
2	Requirements	2
3	5 simple steps to run AlexandrusPS	3
4	Example	7
5	AlexandrusPS applications and functionalities	8
5.1	SUBSTEP 1: Index generation, FASTA header and sequence modification, preparation of files for orthology prediction and quality control	8
5.2	SUBSTEP 2: Orthology prediction by ProteinOrtho	8
5.3	SUBSTEP 3: Selection of the orthology clusters from ProteinOrtho that are suitable for the positive selection analysis	8
5.4	SUBSTEP 4: Calculation of the correct number of cores that will be used for Alexandrus.sh	9
5.5	SUBSTEP 5: For each Orthology group selected in SUBSTEP 3 extract the CDS sequences	9
5.6	SUBSTEP 6: Simplification of the headers of the CDS and amino acid FASTA files	9
5.7	SUBSTEP 7: Peptide alignment performed by PRANK [1]	9
5.8	SUBSTEP 8: Peptide alignment performed by PRANK in nexus format plus phylogenetic tree in nexus format	10
5.9	SUBSTEP 9: Rename and reformat nexus phylogenetic tree of SUBSTEP 8	10
5.10	SUBSTEP 10: Run pal2nal [2]	10
5.11	SUBSTEP 11: Tree labeling according to branches	10
5.12	SUBSTEP 12: Generate configuration files for site-specific models	11
5.13	SUBSTEP 13: Run CodeML for site-specific models	11
5.14	SUBSTEP 14: Quality control of the CodeML output	11
5.15	SUBSTEP 15: Extract information for calculation of LRTs (log ratio tests) for site-specific models	11
5.16	SUBSTEP 16: LRT calculation (log ratio tests) for site-specific models	12
5.17	SUBSTEP 17: Label single species for branch-site analysis.	12
5.18	SUBSTEP 18: Generate configuration files for branch and branch-site models	12
5.19	SUBSTEP 19: Run CodeML for branch and branch-site models	12
5.20	SUBSTEP 20: Extract information for LRT (log ratio tests) calculation for branch and branch-site models	13
5.21	SUBSTEP 21: LRT (log ratio tests) calculation for branch and branch-site models	13
6	Alternative to Docker	13
7	References	15

1 Introduction

AlexandrusPS is a high-throughput user-friendly pipeline designed to simplify the genome-wide positive selection analysis by deploying well-established protocols of CodeML [3]. This can be especially advantageous for researchers with no evolutionary or bioinformatics experience.

AlexandrusPS's main aim is to overcome the technical challenges of a genome-wide positive selection analysis such as i) the execution of an accurate orthology analysis as a precondition for positive selection analysis; ii) preparing and organizing configuration files for CodeML; iii) doing a positive selection analysis on large sets of sequences and iv) generate an output that is easy to interpret including all relevant maximum likelihood (ML) and log ratio test (LRT) results.

The only input data AlexandrusPS needs are the CDS and amino acid sequences of interest. AlexandrusPS provides a simplified output that comprises a table including all relevant results which can be easily extracted for assessment and publication. AlexandrusPS produces and provides all intermediate data such as the results of the ProteinOrtho orthology analysis and the multiple alignments. Default parameters of all steps can be adjusted.

2 Requirements

The easiest way to run AlexandrusPS is to use its Docker image. You can download Docker [here](#).

`docker pull vivianschoonenberg/alexandrusps:0.6` Available tags can be found [here](#).

How to Docker Start an interactive bash shell with the alexandrusps container:

`docker run -it vivianschoonenberg/alexandrusps:0.6` You will be in proper location to run AlexandrusPS. To quit the container, type 'exit'.

To access local files (necessary), you can mount your home or a different folder in the container:

`docker run -rm -mount "type=bind,src=/Users/(id - un),dst = /app/(id -un)" -u (id - u) :(id -g) -it vivianschoonenberg/alexandrusps:0.6` Here, src is the absolute path to the folder you would like to mount. Dst specifies the folder to be mounted in the "app" directory with your username/id. The app folder contains the AlexandrusPS pipeline as well, which is the folder you automatically enter when starting a container from the image (you can move up to the "app" folder using `cd ..`).

You can also use the mounted folder in the container to copy any result or output files to your own local system.

3 5 simple steps to run AlexandrusPS

Sequence name indexing and quality control

Step 1

- For each species that you want to include in the analysis two FASTA files should be generated, one with the amino acid sequences and the other one with correspondent CDS sequences (the same as the amino acid sequences but as CDS sequences). It is crucial that both files have the same number of sequences and that each amino acid sequence and the corresponding CDS sequence have the same header. For example: if you want to analyze 6 different species, you should provide 12 FASTA files (6 '.CDS.fasta' and 6 '.pep.fasta' files), make sure to follow a similar structure as the example data set in the './Example' (Fig. 2N) directory, see Figure 1.

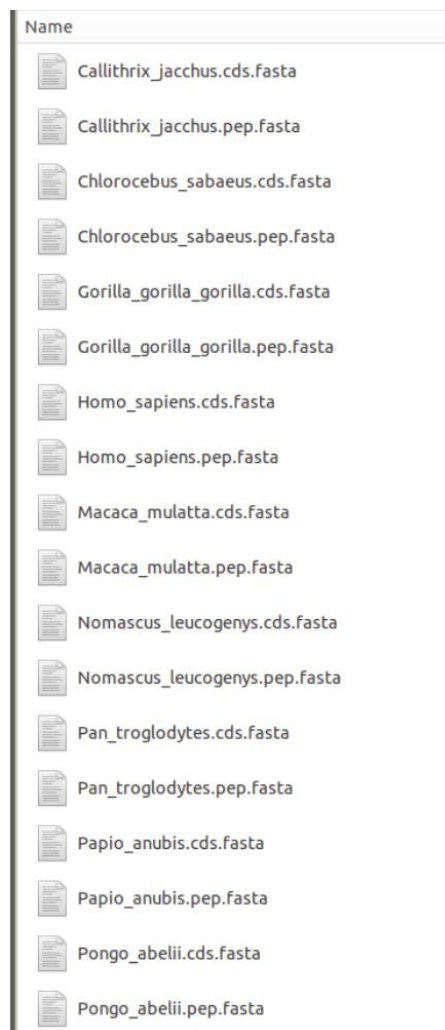


Figure 1: Example sequence files with correct naming

Step 2

- Enter to the main directory of AlexandrusPS (cd './AlexandrusPS') and paste the sequence FASTA files into the directory './Fasta'.

Step 3

- Follow binomial nomenclature rules for naming the FASTA files, this formatting ensures the proper functioning of the pipeline. Here a step by step example for human:

- 1) Find the scientific name for human in binomial nomenclature ("two-term naming system") in which first term is genus or generic name = *Homo* and the second term is the specific name or specific epithet = *sapiens*
- 2) Join the two terms by underline (-) = *Homo.sapiens*
- 3) Add the termination character '.cds.fasta' for the CDS file and '.pep.fasta' for the amino acid files = *Homo.sapiens.cds.fasta* (CDS FASTA file) and *Homo.sapiens.pep.fasta* (amino acid FASTA file).

Two important considerations are:

- i) Both FASTA files need to have the same name, the only difference should be the file extension ('.cds.fasta' and '.pep.fasta').
- ii) AlexandrusPS includes the script `APS1_IndexGenerator_QualityControl.pl` which generates a species name index based on 6 letters from the binomial name - three from the genus (*hom*) and three from the specific epithet (*sap*) - resulting in species name index '*homsap*'. Hence the user should make sure that the file names include only the species name (without special characters besides the mentioned '-') and that the 6 letters do not overlap with the species name index of any other species included in the analysis.

Step 4

- Quality control of your sequences (highly recommended to perform before running `./AlexandrusPS.sh`)
 After you added your sequence FASTA files to the './Fasta' directory (Fig. 2K) and before you run `AlexandrusPS.sh`, we highly recommend to run the script '`Sequences_quality_control_AlexandrusPS.sh`' (Fig. 2A) to check whether your sequence files ('.cds.fasta' and '.pep.fasta') are suitable for positive selection analysis with AlexandrusPS. In case your sequences (either one or both) are not suitable for AlexandrusPS you will find one or two error files ('`Error_missed_sequences.txt`' (Fig. 3A) and/or '`Error_with_Fasta_header.txt`') (Fig. 3B) in the main directory in which you executed '`Sequences_quality_control_AlexandrusPS.sh`' (Fig. 2A). If after running the script none of these files appear it means your sequence files are usable for the analysis. The content of the error files is described and explained in this github repository under the section 'Troubleshooting errors that you may encounter during quality control'. The quality control is performed by the Perl script '`APS1_IndexGenerator_QualityControl.pl`' which is also called and executed by '`Sequences_quality_control_AlexandrusPS.sh`' (Fig. 2A).

Note that this quality control is by default executed by the AlexandrusPS pipeline. The pipeline will continue the analysis with the sequences that pass the quality control even if there are some sequences in '`Error_missed_sequences.txt`' by excluding these from the analysis. It will however interrupt the process if it finds the file '`Error_with_Fasta_header.txt`'.

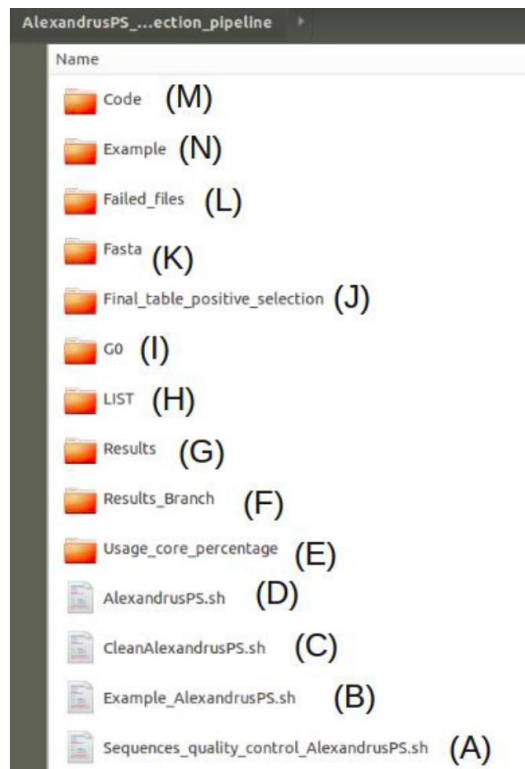


Figure 2: Content of the main directory of AlexandrusPS before execution of AlexandrusPS.sh

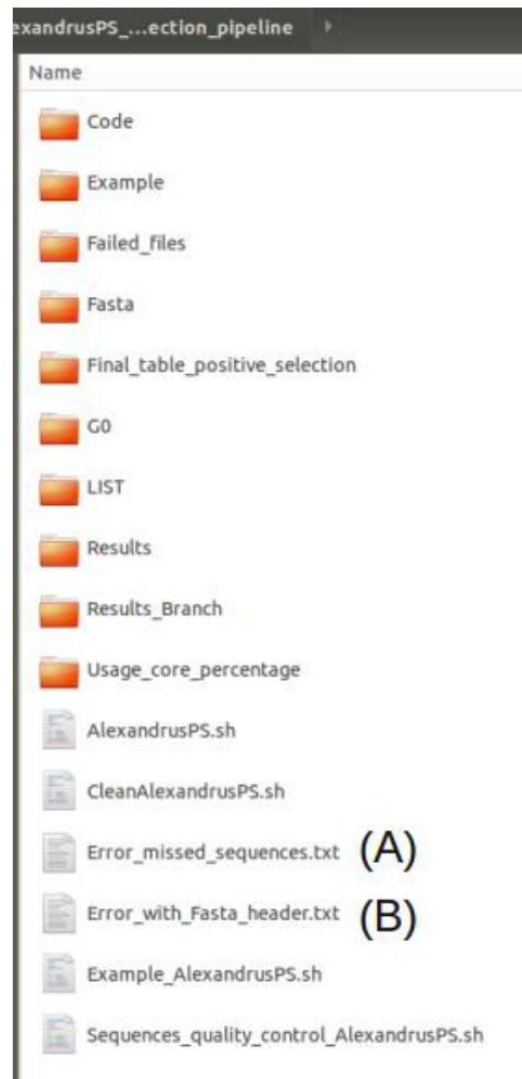


Figure 3: Error files A) Not all amino acids sequences in the '.pep.fasta' file are represented in the '.cds.fasta' file, B) The headers in the '.pep.fasta' and '.cds.fasta' files are different.



Figure 4: Content of the main directory of AlexandrusPS after execution of AlexandrusPS.sh

Troubleshooting errors that you may encounter during quality control In the quality control step AlexandrusPS looks for two main errors in the FASTA files:

- i) Not all amino acids sequences in the '.pep.fasta' file are represented in the '.cds.fasta' file, in which case, the script Sequences_quality_control_AlexandrusPS.sh will generate an error file 'Error_missed_sequences.txt' (Fig. 3A) with all the peptide sequences which could not be found in the '.cds.fasta' file.
- ii) The FASTA file is empty or/and contains empty FASTA entries (header but no sequence) or/and the '.pep.fasta' and the '.cds.fasta' files do not contain the same amount of sequences. In case any of these errors occur it will generate an empty file "Error_with_Fasta_header.txt" (Fig. 3B). If you encounter this error file, we recommend that you re-check the FASTA files and in particular the headers of your FASTA files ('.cds.fasta' and '.pep.fasta'). In general 1) avoid the use of special characters and 2) try to make your headers as short and simple as possible. In case any of these two error files are generated AlexandrusPS will stop execution.

Step 5

- After confirming that no error files were generated in step 4, AlexandrusPS can be executed from the main directory by typing './AlexandrusPS.sh' (Fig. 2D) in terminal.

4 Example

To run the example, navigate to the main directory of the pipeline (Fig. 2) in your terminal and start the analysis by typing './Example_AlexandrusPS.sh' (Fig. 2B).

This executable will transfer the FASTA files from the example directory to the FASTA directory and execute AlexandrusPS.sh with the example dataset provided together with the pipeline.

The output of this example analysis will include the following result: five of the six protein ortho groups

included in the analysis are found to be under positive selection (HLA-DPA1, TLR1, NKG7, CD4, TLR8) and one without positive selection (NUP62CL).

5 AlexandrusPS applications and functionalities

The following explains all the substeps and scripts (in perl or R) that are executed sequentially once AlexandrusPS has been initialized, focusing on:

- 1) Function
- 2) Input files
- 3) Output

5.1 SUBSTEP 1: Index generation, FASTA header and sequence modification, preparation of files for orthology prediction and quality control

Function: Some of the downstream programs of the pipeline struggle with lengthy headers or species names. Such problems are circumvented with the script 'APS1_IndexGenerator_QualityControl.pl' which creates a species name index based on the user-provided binomial name. Using this index the script:

- i) generates FASTA filenames (for .pep.fasta and .cds.fasta) compatible with other downstream scripts used in AlexandrusPS
- ii) adds the index to the headers of the sequences in each FASTA file
- iii) generates a species name index directory enabling the user to retrace the association between the used index and the species' binomial name
- iv) The new headers of the amino acid FASTA file (.cur.pep.fasta) will be used for orthology prediction
- v) compiles the new headers of all species in one file (CompiledSpecies.pep.fasta and CompiledSpecies.cds.fasta) (Fig. 4C) vi) as described before, this SUBSTEP also executes the initial quality control of the sequence files.

Input file: Species_1.pep.fasta and Species_1.cds.fasta

Output: the output files of 'APS1_IndexGenerator_QualityControl.pl' will be located in two new directories created by AlexandrusPS:

- './Curated-Sequences' which will contain the 'CompiledSpecies.pep.fasta' and 'CompiledSpecies.cds.fasta' files (Fig. 4C)
- './Orthology_Prediction' which will contain the '.cur.pep.fasta' files (Fig. 4A).

5.2 SUBSTEP 2: Orthology prediction by ProteinOrtho

Function: Executes ProteinOrtho.

Input file: In SUBSTEP 1 AlexandrusPS.sh generates a list of the cur.pep.fasta files (list_of_pepFiles.txt) in the './Orthology_Prediction' directory

This list is the only argument for the script './Code/APS1_IndexGenerator_QualityControl.pl' (Fig. 2M).

Output: './Orthology_Prediction directory/ProteinOrthoTable.proteinortho' (Fig. 4A)

5.3 SUBSTEP 3: Selection of the orthology clusters from ProteinOrtho that are suitable for the positive selection analysis

Function: Selects ProteinOrtho clusters (orthology groups or OGC) which are suitable for positive selection analysis by the following criteria (produced by 'APS4_OptimalProteinOrthoGroups.pl'):

- 1) OGC encompassing at least three species
- 2) 1-to-1 orthologs (absence of paralogs in any species of the orthologous cluster).

The script extracts the headers of the sequences of the ProteinOrtho clusters which fulfill the requirements for the positive selection analysis, and generates a list with all the ProteinOrtho clusters that will be part of the positive selection analysis.

Input file: Original output table from the ProteinOrtho analysis (ProteinOrthoTable.proteinortho)

Output: Filtered table of Proteinortho table with OGCs which fulfill the requirements ('./Orthology_Prediction/ProteinOrthoTable.proteinortho.fill') (Fig. 4A), the list of orthologous gene cluster IDs (OGC_id) and files with the headers of all proteins from each orthologous gene cluster named by OGC_id located in './LIST/[_OGC_id_].list' (Fig. 3H).

5.4 SUBSTEP 4: Calculation of the correct number of cores that will be used for Alexandrus.sh

Function: Find the number of CPU cores that will be used for the AlexandrusPS positive selection analysis part considering the desired usage percentage provided by the user and leaving 2 free cores for continuing normal usage of the computer, thus avoiding a computer system collapse. The executing script is './Code/APS5_CoreCalculator.pl' (Fig. 2M) and the default usage value is 100

Input file: In the directory './Usage_core_percentage/usage_core_percentage.txt' (Fig. 2E) the user can change the desired usage percent (just the number without the percent symbol - defaults to 100).

Output: './Data/Number_cores.txt.calculated' (Fig. 4B), the number of cores to be used results from the formula: (desired usage percent) * (number of CPU cores available on the computer - 2) / 100
Fig. 5. Contents of directory G0 (Fig. 2I)

5.5 SUBSTEP 5: For each Orthology group selected in SUBSTEP 3 extract the CDS sequences

Function: In this substep the sequences of the orthologs of all relevant orthologous gene clusters are extracted from the sequence FASTA files that were provided by the user. The resulting files are used for the subsequent alignment of the CDS sequences, a crucial step for the positive selection analysis. The executing scripts are './G0/Code/APS7_Extract_Pep_sequences.pl' and './G0/Code/APS8_Extract_Cds_sequences.pl' (Fig. 5A).

Input file: './G0/Orthology_Groups/CompiledSpecies.pep.fasta',
 './G0/Orthology_Groups/CompiledSpecies.cds.fasta'
(prepared in SUBSTEP 1) and './G0/Orthology_Groups/[_OGC_id_]/[_OGC_id_].list' (prepared in SUBSTEP 3) (Fig. 5C).

Output: The FASTA files that will be used for the alignments './G0/Orthology_Groups/[_OGC_id_].list.cds.fasta' and '[_OGC_id_].list.pep.fasta' (Fig. 5C).

5.6 SUBSTEP 6: Simplification of the headers of the CDS and amino acid FASTA files

Function: The script 'APS9_HeaderDictionary_pepCDS.pl', generates new amino acid and CDS FASTA files (.dict.fa) with simplified headers leaving just the species name index (see SUBSTEP 2) followed by [_OGC_id_] and a number assigned in the OGC's alignment. It also generates a dictionary associating the new with the old headers (original headers provided by the user), for both amino acid and CDS FASTA files (.fasta.dict).

Input: FASTA files generated in SUBSTEP 5 ([_OGC_id_].list.cds.fasta and [_OGC_id_].list.pep.fasta)

Output: 1) Dictionaries: './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict' and './G0/Orthology_Groups/[_OGC_id_].list.cds.fasta.dict'

2) FASTA files: './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict.fa' and './G0/Orthology_Groups/[_OGC_id_].list.cds.fasta.dict.fa' (Fig. 5C)

5.7 SUBSTEP 7: Peptide alignment performed by PRANK [1]

Function: CodeML is based on codon alignments, for that reason peptide alignment of all proteins in the respective orthologous groups is performed using PRANK.

Input: './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict.fa' (prepared in SUBSTEP 6) (Fig. 5C)

Output: Alignment files './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict.fa.best.fas' (Fig. 5C)

5.8 SUBSTEP 8: Peptide alignment performed by PRANK in nexus format plus phylogenetic tree in nexus format

Function: CodeML needs peptide alignment information and phylogenetic gene trees in nexus format. The executing program PRANK provides these formats.

Input: './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict.fa' (Fig. 5C)

Output: './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict.fa.best.nex' (Fig. 5C)

5.9 SUBSTEP 9: Rename and reformat nexus phylogenetic tree of SUBSTEP 8

Function: CODEML requires a phylogenetic tree with headers of the FASTA file. As PRANK does not provide this, the script './G0/Code/APS10_CleanNex_nex.pl' (Fig. 5A) takes the nexus alignment ('.best.nex') of SUBSTEP 9, extracts the phylogenetic tree ('.best.nex.cl.nex') and the numeration of each species and the association with the header from the alignment ('.best.nex.dict') and replaces the automated numeration generated by PRANK with the header of the FASTA file in the phylogenetic tree ('.best.nex.cl.head.nex'). The script also changes nexus to dnd format making this compatible with other downstream steps ('.best.nex.cl.head.dnd').

Input: nexus file of SUBSTEP 9 './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict.fa.best.nex' (Fig. 5C)

Output:

- 1) Phylogenetic tree from PRANK in nexus format:
 './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict.fa.best.nex.cl.nex' (Fig. 5C)
- 2) Dictionary associating the numeration generated by PRANK with the header of the amino acid FASTA file
 './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict.fa.best.nex.dict' (Fig. 5C)
- 3) Nexus tree with species names: './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict.fa.best.nex.cl.head.nex' (Fig. 5C)
- 4) Format change from nexus to dnd
 './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict.fa.best.nex.cl.head.dnd' (Fig. 5C)

5.10 SUBSTEP 10: Run pal2nal [2]

Function: As CodeML is a codon-based model the multiple sequence alignment of proteins ('.pep.fasta.dict.fa.best.fas') and the corresponding CDS (.list.cds.fasta.dict.fa) sequences need to be converted into a codon alignment (.codonalign.fasta). This is achieved using pal2nal.

Input:

1) Multiple sequence alignments of proteins generated in SUBSTEP 8
 './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict.fa.best.fas' 2) CDS sequences generated in SUBSTEP 7 './G0/Orthology_Groups/[_OGC_id_].list.cds.fasta.dict.fa'

Output: CDS codon alignment:

'./G0/Orthology_Groups/[_OGC_id_].codonalign.fasta' (Fig. 5C)

5.11 SUBSTEP 11: Tree labeling according to branches

Function: In order to enable branch analysis any tree needs to be provided with the corresponding labels.

Input: './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict.fa.best.nex.cl.head.dnd' (Fig. 5C)

Output: './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict.fa.best.nex.cl.head.dnd.GenTree.nex' (Fig. 5C)

5.12 SUBSTEP 12: Generate configuration files for site-specific models

Function: Generates configuration files to run site-specific model analyses that fit seven codon substitution models: M0 ('./G0/Code/APS12_CreateCtl_ParameterDefParPG_M0.pl') (Fig. 5A), M1a, M2a, M3, M7 ('./G0/Code/APS13_CreateCtl_ParameterDefParPG_SM.pl') (Fig. 5A), M8 and M8a ('./G0/Code/APS14_CreateCtl_ParameterDefParPG_SM8.pl') (Fig. 5A). It uses the configuration file './Data/Parameter.codeml.M0.txt' (for APS12), './Data/Parameter.codeml.SM.txt' (for APS13) or './G0/Data/Parameter.codeml.SM8.txt' (for APS14) (Fig. 5B) (these configuration files can be modified by the user) and a default configuration file ('./G0/Data/Default.par.txt') that fills any lacking information in the executed configuration file ('./G0/Data/Parameter.codeml.M0.txt') (Fig. 5B).

Input: './G0/Data/Parameter.codeml.M0.txt' (for APS12) or './G0/Data/Parameter.codeml.SM.txt' (for APS13) or './G0/Data/Parameter.codeml.SM8.txt' (for APS14) and './G0/Data/Default.par.txt' (Fig. 5B)

Output: Configuration files './G0/Orthology_Groups/codeml_OGC_id_.M0.ctl', './G0/Orthology_Groups/codeml_OGC_id_.sm8.ctl' and './G0/Orthology_Groups/codeml_OGC_id_.sm.ctl' (Fig. 5C)

5.13 SUBSTEP 13: Run CodeML for site-specific models

Function: Run CodeML using the configuration files (.ctl) generated in SUBSTEP 12.

Input: Configuration files './G0/Orthology_Groups/codeml[_OGC_id_.]M0.ctl', './G0/Orthology_Groups/codeml[_OGC_id_.]sm8.ctl' and './G0/Orthology_Groups/codeml[_OGC_id_.]sm.ctl' (Fig. 5C)

Output: configuration files './G0/Orthology_Groups/codeml[_OGC_id_.]M0.mlc', './G0/Orthology_Groups/codeml[_OGC_id_.]sm8.mlc' and './G0/Orthology_Groups/codeml[_OGC_id_.]sm.mlc' (Fig. 5C)

5.14 SUBSTEP14: Quality control of the CodeML output

Function: In cases when CodeML cannot perform the analysis, the output from CodeML does not contain the information necessary for LRT (log ratio tests) calculation. To exclude these instances from the global results output they are marked for exclusion.

Input: All CodeML output files './G0/Orthology_Groups/codeml[_OGC_id_.]M0.mlc', './G0/Orthology_Groups/codeml[_OGC_id_.]sm8.mlc' and './G0/Orthology_Groups/codeml[_OGC_id_.]sm.mlc' (Fig. 5C)

Output: In case of missing data will create a file called ErrorInTable.txt, which is used to condition SUBSTEP 15 Figure 6. Outputs files in ./Final_table_positive_selection (Fig. 2J) after AlexandrusPS.sh finishes

5.15 SUBSTEP 15: Extract information for calculation of LRTs (log ratio tests) for site-specific models

Function: The output files of CodeML (.mlc files) which include the site-specific models performed in SUBSTEP 13 need parsing to extract the information needed for LRT calculation. This task is performed by the script './G0/Code/APS16_ExtractLRTandNP_positiveSelection.pl' (Fig. 5A) which extracts: log likelihood (lnL), the number of parameters (np) for each model, omega for M0, M8, p0 and p1 for M1 and M8, w0 and w1 for M1 and the positive selection sites (PSS, aminoacid under selection) for all the models.

Input: All the output files of CodeML './G0/Orthology_Groups/codeml[_OGC_id_.]M0.mlc', './G0/Orthology_Groups/codeml[_OGC_id_.]sm8.mlc' and './G0/Orthology_Groups/codeml[_OGC_id_.]sm.mlc' (Fig. 5C)

Output: If all the models have complete information the table

'./Final_table_positive_selection/PositiveSelectionTable.txt' (Fig. 6F) will be filled with data. If important CodeML values such as the likelihood (lnL) and/or the number of parameters (np) are missing, the table './Failed_files/FailedPositiveSelectionTable.txt' (Fig. 2L) will be filled with any available information and absent data replaced with NAs.

5.16 SUBSTEP 16: LRT calculation (log ratio tests) for site-specific models

Function: LRT calculation and FDR correction based on the data in table './Final_table_positive_selection/PositiveSelectionTable.txt' (Fig. 6F) is performed by R script './G0/Code/APS18_Calculate_LTR.R.' (Fig. 5A).

Input: Table './Final_table_positive_selection/PositiveSelectionTable.txt' (Fig. 6F) .

Output: A table including only the genes under positive selection at the site-specific level './Final_table_positive_selection/GenesUnderPositiveSelection.txt' (Fig. 6G). All intermediate files (from SUBSTEP 5 to 16) of all genes that do not show signals of positive selection will be compressed in './Results/[_OGC_id_].tar.gz' (Fig. 2G).

5.17 SUBSTEP 17: Label single species for branch-site analysis.

Function: In order to assess positive selection for individual branches of the phylogeny, this substep generates an equal number of trees as species in the orthology group - for each a different species is labeled as the foreground branch, leaving the rest of the species as background branches. This is performed by the script './G0/Code/APS19_TreeGeneratorCombinator.pl' (Fig. 5A).

Input: Table './G0/Orthology_Groups/[_OGC_id_].list.pep.fasta.dict.fa.best.nex.cl.head.dnd.GenTree.nex' from SUBSTEP 11 (Fig. 5C)

Output: Labeled tree for each species in the respective orthology group './G0/Orthology_Groups/[species_name_index][_OGC_id_].BranchAnalyTree' and a list of trees TreeList.txt (Fig. 5C)

5.18 SUBSTEP 18: Generate configuration files for branch and branch-site models

Function: Generates configuration files to run branch-site model analyses that fit seven codon substitution models: M0 ('./G0/Code/APS20_CreateCtl_ParameterDefParPG_BSM0.pl'), H0 ('./G0/Code/APS21_CreateCtl_ParameterDefParPG_BSM0H0.pl') (Fig. 5A), and H1 ('./G0/Code/APS22_CreateCtl_ParameterDefParPG_BSM0H1.pl') (Fig. 5A), using the configuration file './Data/Parameter_codeml_M0BS.txt' (Fig. 5A) (for APS20) or './G0/Data/Parameter_codeml_M2BSH0.txt' (for APS21) or './G0/Data/Parameter_codeml_M2BSH1.txt' (for APS22) (these files can be modified by the user) and a default configuration file './G0/Data/Default_par.txt' that fills any gap in the executed CodeML configuration files (Fig. 5B).

Input: './G0/Data/Parameter_codeml_M0BS.txt' (for APS20) or './G0/Data/Parameter_codeml_M2BSH0.txt' (for APS21) or './G0/Data/Parameter_codeml_M2BSH1.txt' (for APS22) and './G0/Data/Default_par.txt.' (Fig. 5B)

Output: Configuration files './G0/Orthology_Groups/codeml[species_name_index][_OGC_id_].bsm0.ctl', './G0/Orthology_Groups/codeml[species_name_index][_OGC_id_].bsm0h0.ctl' and './G0/Orthology_Groups/codeml[species_name_index][_OGC_id_].bsm0h1.ctl' (Fig. 5C)

5.19 SUBSTEP 19: Run CodeML for branch and branch-site models

Function: Run CodeML with the configuration files (.ctl) generated in SUBSTEP 18.

Input: Configuration files './G0/Orthology_Groups/codeml[species_name_index][_OGC_id_].bsm0.ctl', './G0/Orthology_Groups/codeml[species_name_index][_OGC_id_].bsm0h0.ctl' and './G0/Orthology_Groups/codeml[species_name_index][_OGC_id_].bsm0h1.ctl' (Fig. 5C)

Output: CodeML output files './G0/Orthology_Groups/codeml[species_name_index][_OGC_id_].bsm0.mlc',

‘./G0/Orthology_Groups/codeml[species_name_index][_OGC_id_].bsm0h0.mlc’ and
‘./G0/Orthology_Groups/codeml[species_name_index][_OGC_id_].bsm0h1.mlc’ (Fig. 5C).

5.20 SUBSTEP 20: Extract information for LRT (log ratio tests) calculation for branch and branch-site models

Function: The CodeML output files (.mlc files) of the branch and branch-site model analyses performed in SUBSTEP 19 need to be parsed for the information needed for LRT calculation. This is performed by the script ‘./G0/Code/APS23_ExtractLRTandNP_positiveSelectionBranchSite.pl’ (Fig. 5A) which extracts: likelihood (lnL) and number of parameters (np) for each model.

Input: Full CodeML output ‘./G0/Orthology_Groups/codeml[species_name_index][_OGC_id_].bsm0.mlc’, ‘./G0/Orthology_Groups/codeml[species_name_index][_OGC_id_].bsm0h0.mlc’ and ‘./G0/Orthology_Groups/codeml[species_name_index][_OGC_id_].bsm0h1.mlc’ (Fig. 5C)

Output: Table including all relevant data for LRT calculation
‘./Final_table_positive_selection/Branch_models_BranchSite_models_Table.txt’ (Fig. 6B)

5.21 SUBSTEP 21: LRT (log ratio tests) calculation for branch and branch-site models

Function: LRT calculation and FDR correction based on the data in table ‘./Final_table_positive_selection/Branch_models_BranchSite_models_Table.txt’ (Fig. 6B) performed by the R script ‘./G0/Code/APS23_BranchSiteAnalysis.R’. (Fig. 5A).

Input: Table including all relevant data for LRT calculation from SUBSTEP 20
‘./Final_table_positive_selection/Branch_models_BranchSite_models_Table.txt’.

Output: Table including only genes under positive selection at the branch and branch-site level ‘./Final_table_positive_selection/Branch_model.txt’ (Fig. 6A) and ‘./Final_table_positive_selection/Branch_site_model.txt’ (Fig. 6C).
The intermediate files (from SUBSTEP 5 to 21) will be compressed in (‘./Results_Branch/[_OGC_id_]bs.tar.gz’) (Fig. 2F).

6 Alternative to Docker

AlexandrusPS was devised to run without any previous installation given the docker container. Nevertheless, the user is given the choice to install all the necessary programs and modules independently. Perl
Perl 5: <https://www.perl.org/> The following perl modules are required and can be installed them using cpan:

- Data::Dumper
- List::MoreUtils qw(uniq)
- Array::Utils qw(:all)
- String::ShellQuote qw(shell_quote)
- List::Util
- POSIX

R version 4.0.5

R: <https://www.r-project.org/> The following libraries are necessary:

- dplyr
- ggplot2
- caret
- reshape2
- ggpubr

- RColorBrewer
- stringr
- viridis
- Rstatix

Protein orthology search ProteinOrtho (<https://www.bioinf.uni-leipzig.de/Software/proteinortho/>) v6.06
Aligners PRANK multiple sequence aligner (<http://wasabiapp.org/software/prank/>) v.170427 PAL2NAL
<http://www.bork.embl.de/pal2nal/Download> v14 PAML The PAML software package includes CodeML
(<http://abacus.gene.ucl.ac.uk/software/paml.html>) - v4.8a or v4.7

7 References

References

- [1] Ari Löytynoja. Phylogeny-aware alignment with prank. *Multiple sequence alignment methods*, pages 155–170, 2014.
- [2] Mikita Suyama, David Torrents, and Peer Bork. Pal2nal: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research*, 34(suppl.2):W609–W612, 2006.
- [3] Ziheng Yang. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591, 2007.

5 Discussion

As part of the characterization of the *C. elegans* telomere-binding proteins TEPB-1 and TEPB-2 (Article I - [110]) we tried to strengthen our finding by using evolutionary analyses using gene annotations of other nematode species from WormBase. In the course of these efforts, we realized that although the data available for *C. elegans* is of very high quality, annotations of many other species suffer from incompleteness and inaccuracy represented as partial, falsely merged, or missing gene models. This gap is detrimental to the accuracy of evolutionary comparisons and will lead to misinterpretations in many cases. As this problem occurred in many of the other projects in the lab, we tried to tackle the annotation problem using a genome assembly-independent method. Here, Proteotranscriptomics seemed to provide a feasible methodology as proteomics as well as RNA-sequencing is easily attainable. When implementing this technique combined with a machine learning-assisted quality control on an interesting set of 12 nematodes we could show that indeed the produced annotations exhibit exceptionally high quality comparable to that of well-established model organisms such as *C. elegans* (Article II - [1]). The study included species without available genome assemblies, highlighting the great potential of this method for any species of interest. To further exemplify the power of this dataset we set forth to perform a genome-wide positive selection analysis across all 12 species. To solve all related obstacles in performing such analyses across all possible orthology groups, we developed a pipeline that facilitates such analyses in an automated fashion without the need for high-performance computing systems. We expect this pipeline to be of value to many other labs and hence want to provide the workflow in a Docker image to the broader scientific community (Article III -in preparation).

The evolutionary analyses performed in the framework of Article II [1] provided a myriad of highly interesting clues into evolutionary dynamics in the different subgroups of the nematode set included. Below some of the major findings are discussed in greater detail.

5.1 Biases related to the reproductive mode in nematodes

Nematodes interestingly have very different models of reproduction. The species included in our study (Article II - [1]) included *C. briggsa*, *O. tipulae*, *P. pacificus* and *C. elegans*, which are androdioecious (primarily selfing) and 8 other species that are gonochoristic (mating). The reproductive mode in nematodes has been demonstrated to be a critical factor in

the shaping of genetic diversity and effective population size. Androdioecy, marked by homozygosity and reduced effective recombination, can result in increased mutational biases and a buildup of deleterious mutations, leading to a shift away from selection and towards genetic drift. On other hand, gonochoristic species are expected to have high heterozygosity[7]. These characteristics are expected to affect the transcriptome assembly process and the analysis of positive selection (dN/dS). On the one hand, individuals with high levels of heterozygosity (gonochoristic) may pose challenges in transcriptome assembly. On the other hand, the trend towards genetic drift (androdioecious) should be evident when evaluating dN/dS.

To investigate the impact of reproductive mode on transcriptome assembly and positive selection analysis, the studied species were divided into two groups based on their reproductive modes. The first measure involved an evaluation of the impact of reproductive mode on the transcriptome assembly process. To assess this, we calculated the ratio of the number of assembled contigs to the number of reads utilized for the assembly. Our results showed that despite the well-known impact of individual heterozygosity on genome assemblies, our transcriptome assembly was not affected by this issue. Furthermore, our analysis found no significant differences in the number of high-quality assembled transcripts across species that have distinct reproductive modes (Article II -Supplemental Fig. S9B). The second measure involved the investigation of the effect of reproductive mode on the detection of positive selection. Our results demonstrate the absence of significant trends in terminal branch average dN/dS values dependent on the reproductive mode in nematodes (Article II -Supplemental Fig. S9C). Previous studies have indicated that the impact of reproductive mode on genetic diversity and effective population size can be understood through its association with intergenic and intron regions. In particular, the androdioecious reproductive mode has been shown to facilitate intron evolution, leading to a greater impact of insertion/deletion mutational biases on intron size [7]. This suggests that the effect of the reproductive mode may be more pronounced in non-coding regions than in coding regions. This provides additional evidence to the absence of any biases related to the reproductive mode in our data.

5.2 Enrichments of networks related to cell division and spindle organization in the *Caenorhabditis* genus

Characterizing the ortholog groups with positive selection for *Caenorhabditis*, we found enrichments of networks related to cell division and spindle organization. While it is known that species of the *Caenorhabditis* genus

have unique spindle formation mechanisms [111], the assembly and disassembly of the involved protein complexes is 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, the functional homologs of human CDK5RAP2/Cnn [112] that belong to the main components of the pericentriolar material (PCM) suggested to be involved in the conformation of the mitotic spindle [113] [114], ROD, involved in the chromosomal segregation during cell division [115], the microtubule-stabilizing and nucleation-promoting factor TPXL-1 playing an important role in the regulator of spindle assembly, a paralog of the microtubule destabilizer KLP-7, KPL19 [116] [117] and LET-92 that is involved in the disassembly of SPD-5. We show that among nematodes these genes are indeed unique and common for the *Caenorhabditis* genus. These results suggest that other genes in this group might also have functions in spindle organization.

5.3 *C. elegans* genes with positive selection evidence related to muscle function

Our positive selection analysis revealed that *C. elegans* has a higher number of positively selected genes related to muscle functions. To understand the evolutionary impact at the dN/dS level, one would need to take into account factors such as generation time and the effective population size. A loss of neutral genetic polymorphism is theoretically expected in androdioecious organisms when compared to gonochoristic. One reason, as described above, derives from the decrease of the effective population size in purely androdioecious species which generates a genetic bottleneck. Previous studies have demonstrated that bottlenecks are less frequent in hermaphroditic animals [118]. In order to understand the enrichment of muscle-related proteins in our study, we sought to evaluate the life history of *Caenorhabditis elegans* as a model organism in laboratories. The laboratory strain of *C. elegans* has been propagated for approximately 70 years in a controlled environment with constant temperature, light, humidity, and unlimited food, which differs substantially from the conditions experienced by the wild strain. Parque et al. [119] revealed that when placed in microfluidic chambers that mimic complex environments such as soil, the wild strain of *C. elegans* exhibited a new mode of locomotion that combined the fast gait of swimming with the more efficient movements of crawling. This mode was distinct from the smooth surface movement observed on agar plates. Moreover, Gómez-Marín et al. [120] found that wild-type *C. elegans* displayed more ordered locomotion than laboratory

reference strains N2 and N2 mutant. These results suggest that the enrichment of muscle-related functions in *C. elegans* may reflect adaptation to movement on two-dimensional agar plates, rather than a general constraint on population diversity.

5.4 Bias on positive selection detection with high divergent species

In our evaluation of the distribution of branch-site-specific positive signals in the species within ProteinOrtho groups, we observed that these signals were unevenly distributed across species subsets, declining as evolutionary distance increased. Our analysis found 47.9% of the signals in *Caenorhabditis* (1672 orthology groups), 29.3% in *Eurhabditis* (1507 groups), and 16.5% in *Rhabditida* (897 groups) (Article II -Fig. 5C and 5D). The majority of the genes in our study showed evidence of genetic drift, with 45.3% of orthology groups in *Caenorhabditis*, 56.4% in *Eurhabditis*, and 63.2% in *Rhabditida* (Fig. 5C). It is not yet clear whether this pattern is a result of our analysis or a typical trend.

The detection of positive selection is limited by the evolutionary distances between species, with the degree of divergence between species proportional to the sequence identity of aligned orthologous sequences. As evolutionary distances increase, the sequence identity decreases, making it increasingly difficult to accurately detect positive selection. This tendency has been previously observed in *Rhabditida* [1]. The accuracy of CodeML analysis for detecting positive selection can be compromised by the presence of insertions and deletions, which can introduce wrong inferences. The decline in detection of positive selection signals with increasing evolutionary distances is likely due to the decrease in alignment quality as protein similarity decreases [104] [72]. This has to be taken into account and while for close species most of the positive selection will be discovered, for species with large divergence we expect only very significant signals to be detected.

6 Conclusions

The results presented in this thesis demonstrate the efficacy of integrating high-throughput experimental data, such as RNA-seq and peptide evidence, to facilitate accurate protein-coding gene annotation. The utilization of proteotranscriptomics methodology leads to highly valid gene prediction even in species without a reference genome, achieving qualities comparable to well-studied organisms such as *C. elegans*.

Accurate gene annotations are crucial in conducting any evolutionary analyses such as positive selection. To enable the genome-wide positive selection detection the study included the development of the **AlexandrusPS** pipeline, which implements standard **CodeML** protocols and aims to avoid biases in positive selection identification. The pipeline will be valuable for the broader scientific community and hence will be made publicly available as a Docker image to enable easy application even for researchers with limited bioinformatics expertise and computational resources.

The combination of high-quality experimental data and appropriate positive selection analysis allowed for a comprehensive evolutionary analysis in nematodes, extending the understanding gained from decades of research on *C. elegans* to a diverse range of nematode species with different life histories, modes of reproduction, and habitats. This analysis sheds light on how nematode species have evolved to better adapt to their environments through changes in genes involved in stress response, detoxification, metabolism, reproduction, and development.

In conclusion, the use of proteotranscriptomics results in highly reliable and experimentally validated gene annotations without the need for elaborated genome assembly. These annotations have the potential to advance evolutionary studies, including the analysis of positive selection and phylogeny providing insights into the genomic adaptations of nematode species to their environments. This study underscores the importance and impact of large data sets in evolutionary analyses and serves as a valuable foundation for future evolutionary research.

7 Acknowledgements

8 References

References

- [1] Alejandro Ceron-Noriega, Miguel V Almeida, Michal Levin, and Falk Butter. Nematode gene annotation by machine-learning-assisted proteotranscriptomics enables proteome-wide evolutionary analysis. *Genome Research*, 2023.
- [2] Titus Kaletta and Michael O Hengartner. Finding function in novel targets: *C. elegans* as a model organism. *Nature reviews Drug discovery*, 5(5):387–399, 2006.
- [3] Gregory D Plowman, Sucha Sudarsanam, Jonathan Bingham, David Whyte, and Tony Hunter. The protein kinases of *caenorhabditis elegans*: a model for signal transduction in multicellular organisms. *Proceedings of the National Academy of Sciences*, 96(24):13603–13610, 1999.
- [4] Karin Kiontke, Nicholas P Gavin, Yevgeniy Raynes, Casey Roehrig, Fabio Piano, and David HA Fitch. *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proceedings of the National Academy of Sciences*, 101(24):9003–9008, 2004.
- [5] Ralf J Sommer. Evolution of development in nematodes related to *c. elegans*. *WormBook: The Online Review of C. elegans Biology [Internet]*, 2005.
- [6] James H Thomas. Genome evolution in *caenorhabditis*. *Briefings in Functional Genomics and Proteomics*, 7(3):211–216, 2008.
- [7] Soochin Cho, Suk-Won Jin, Adam Cohen, and Ronald E Ellis. A phylogeny of *caenorhabditis* reveals frequent loss of introns during nematode evolution. *Genome Research*, 14(7):1207–1220, 2004.
- [8] Giobbe Forni, Angelo Alberto Ruggieri, Giovanni Piccinini, and Andrea Luchetti. Base: A novel workflow to integrate nonubiquitous genes in comparative genomics analyses for selection. *Ecology and Evolution*, 11(19):13029–13035, 2021.
- [9] Jeb Gaudet and James D McGhee. Recent advances in understanding the molecular mechanisms regulating *c. elegans* transcription. *Developmental dynamics: an official publication of the American Association of Anatomists*, 239(5):1388–1404, 2010.

- [10] Paul W Sternberg and H Robert Horvitz. Postembryonic nongonadal cell lineages of the nematode *panagrellus redivivus*: description and comparison with those of *caenorhabditis elegans*. *Developmental biology*, 93(1):181–205, 1982.
- [11] Todd W Harris, Igor Antoshechkin, Tamberlyn Bieri, Darin Blasiar, Juancarlos Chan, Wen J Chen, Norie De La Cruz, Paul Davis, Margaret Duesbury, Ruihua Fang, et al. Wormbase: a comprehensive resource for nematode research. *Nucleic acids research*, 38(suppl_1):D463–D467, 2010.
- [12] Maria Anisimova and Carolin Kosiol. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Molecular biology and evolution*, 26(2):255–271, 2009.
- [13] Raphael Steffen, Lynn Ogoniak, Norbert Grundmann, Anna Pawluchin, Oliver Soehnlein, and Jürgen Schmitz. papaml: An improved computational tool to explore selection pressure on protein-coding sequences. *Genes*, 13(6):1090, 2022.
- [14] Renato Fani and Marco Fondi. Origin and evolution of metabolic pathways. *Physics of Life Reviews*, 6(1):23–52, 2009.
- [15] Jeffrey H Chuang and Hao Li. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS biology*, 2(2):e29, 2004.
- [16] Adi Stern, Adi Doron-Faigenboim, Elana Erez, Eric Martz, Eran Bacharach, and Tal Pupko. Selecton 2007: advanced models for detecting positive and purifying selection using a bayesian inference approach. *Nucleic acids research*, 35(suppl_2):W506–W511, 2007.
- [17] Luba M Pardo, Ian MacKay, Ben Oostra, Cornelia M van Duijn, and Yurii S Aulchenko. The effect of genetic drift in a young genetically isolated population. *Annals of human genetics*, 69(3):288–295, 2005.
- [18] Alexander Stark, Michael F Lin, Pouya Kheradpour, Jakob S Pedersen, Leopold Parts, Joseph W Carlson, Madeline A Crosby, Matthew D Rasmussen, Sushmita Roy, Ameya N Deoras, et al. Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature*, 450(7167):219–232, 2007.
- [19] Claudio Casola and Matthew W Hahn. Gene conversion among paralogs results in moderate false detection of positive selection using likelihood methods. *Journal of molecular evolution*, 68(6):679–687, 2009.

- [20] Fernando Castro-Chavez. The rules of variation: amino acid exchange according to the rotating circular genetic code. *Journal of theoretical biology*, 264(3):711–721, 2010.
- [21] Juan I Montoya-Burgos. Patterns of positive selection and neutral evolution in the protein-coding genes of tetraodon and takifugu. *PLoS One*, 6(9):e24800, 2011.
- [22] Amy Egan, Anup Mahurkar, Jonathan Crabtree, Jonathan H Badger, Jane M Carlton, and Joana C Silva. Idea: interactive display for evolutionary analyses. *BMC bioinformatics*, 9(1):1–9, 2008.
- [23] Willie J Swanson. Adaptive evolution of genes and gene families. *Current opinion in genetics & development*, 13(6):617–622, 2003.
- [24] Emanuel Maldonado, Daniela Almeida, Tibisay Escalona, Imran Khan, Vitor Vasconcelos, and Agostinho Antunes. Lmap: lightweight multigene analyses in paml. *BMC bioinformatics*, 17(1):1–11, 2016.
- [25] Joao L Reis-Cunha, Hugo O Valdivia, and Daniella Castanheira Bartholomeu. Gene and chromosomal copy number variations as an adaptive mechanism towards a parasitic lifestyle in trypanosomatids. *Current genomics*, 19(2):87–97, 2018.
- [26] Steven Weaver, Stephen D Shank, Stephanie J Spielman, Michael Li, Spencer V Muse, and Sergei L Kosakovsky Pond. Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. *Molecular biology and evolution*, 35(3):773–777, 2018.
- [27] Klaus-Peter Koepfli, Benedict Paten, Genome 10K Community of Scientists, and Stephen J O’Brien. The genome 10k project: a way forward. *Annu. Rev. Anim. Biosci.*, 3(1):57–111, 2015.
- [28] Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F Lin, Brian J Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, 2011.
- [29] Ryan K Schott, Daniel Gow, and Belinda SW Chang. Blastphyme: a toolkit for rapid generation and analysis of protein-coding sequence datasets. *bioRxiv*, page 059881, 2019.
- [30] Laura G Macías, Eladio Barrio, and Christina Toft. Gwidecodeml: a python package for testing evolutionary hypotheses at the genome-wide level. *G3: Genes, Genomes, Genetics*, 10(12):4369–4372, 2020.

- [31] Marc Tollis, Elizabeth D Hutchins, Jessica Stapley, Shawn M Rupp, Walter L Eckalbar, Inbar Maayan, Eris Lasku, Carlos R Infante, Stuart R Dennis, Joel A Robertson, et al. Comparative genomics reveals accelerated evolution in conserved pathways during the diversification of anole lizards. *Genome Biology and Evolution*, 10(2):489–506, 2018.
- [32] Jessica Alföldi and Kerstin Lindblad-Toh. Comparative genomics as a tool to understand evolution and disease. *Genome research*, 23(7):1063–1068, 2013.
- [33] Philipp Khaitovich, Ines Hellmann, Wolfgang Enard, Katja Nowick, Marcus Leinweber, Henriette Franz, Gunter Weiss, Michael Lachmann, and Svante Paabo. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, 309(5742):1850–1854, 2005.
- [34] Ake Liu, Funan He, Libing Shen, Ruixiang Liu, Zhijun Wang, and Jingqi Zhou. Convergent degeneration of olfactory receptor gene repertoires in marine mammals. *BMC genomics*, 20(1):1–14, 2019.
- [35] Maxime Policarpo, Julien Fumey, Philippe Lafargeas, Delphine Naquin, Claude Thermes, Magali Naville, Corentin Dechaud, Jean-Nicolas Volff, Cedric Cabau, Christophe Klopp, et al. Contrasting gene decay in subterranean vertebrates: insights from cavefishes and fossorial mammals. *Molecular biology and evolution*, 38(2):589–605, 2021.
- [36] Qipian Chen, Hao Yang, Xiao Feng, Qingjian Chen, Suhua Shi, Chung-I Wu, and Ziwen He. Two decades of suspect evidence for adaptive molecular evolution—negative selection confounding positive-selection signals. *National Science Review*, 9(5):nwab217, 2022.
- [37] Tanja Schwander, Romain Libbrecht, and Laurent Keller. Supergenes and complex phenotypes. *Current Biology*, 24(7):R288–R294, 2014.
- [38] Joshua A Shapiro, Wei Huang, Chenhui Zhang, Melissa J Hubisz, Jian Lu, David A Turissini, Shu Fang, Hurng-Yi Wang, Richard R Hudson, Rasmus Nielsen, et al. Adaptive genic evolution in the drosophila genomes. *Proceedings of the National Academy of Sciences*, 104(7):2271–2276, 2007.
- [39] Stephanie J Spielman, Suyang Wan, and Claus O Wilke. A comparison of one-rate and two-rate inference frameworks for site-specific dn/ds estimation. *Genetics*, 204(2):499–511, 2016.

- [40] Nancy Manchanda, John L Portwood, Margaret R Woodhouse, Arun S Seetharam, Carolyn J Lawrence-Dill, Carson M Andorf, and Matthew B Hufford. Genomeqc: a quality assessment tool for genome assemblies and gene structure annotations. *BMC genomics*, 21(1):1–9, 2020.
- [41] Arang Rhie, Shane A McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, William Chow, Arkarachai Fungtammasan, Juwan Kim, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592(7856):737–746, 2021.
- [42] Liliana Florea, Alexander Suvorov, Theodore S Kalbfleisch, and Steven L Salzberg. Genome assembly has a major impact on gene content: a comparison of annotation in two bos taurus assemblies. *PLoS One*, 6(6):e21400, 2011.
- [43] Michal Levin and Falk Butter. Proteotranscriptomics-a facilitator in omics research. *Computational and Structural Biotechnology Journal*, 2022.
- [44] Steven L Salzberg. Next-generation genome annotation: we still struggle to get it right, 2019.
- [45] Steven L Salzberg. Genome re-annotation: a wiki solution? *Genome biology*, 8(1):1–5, 2007.
- [46] William Klimke, Claire O’Donovan, Owen White, J Rodney Brister, Karen Clark, Boris Fedorov, Ilene Mizrahi, Kim D Pruitt, and Tatiana Tatusova. Solving the problem: genome annotation standards before the data deluge. *Standards in genomic sciences*, 5:168–193, 2011.
- [47] Peter Bakke, Nick Carney, Will DeLoache, Mary Gearing, Kjeld Ingvorsen, Matt Lotz, Jay McNair, Pallavi Penumetcha, Samantha Simpson, Laura Voss, et al. Evaluation of three automated genome annotations for halorhabdus utahensis. *PloS one*, 4(7):e6291, 2009.
- [48] Jiao Ma, Alan Saghatelian, and Maxim Nikolaievich Shokhirev. The influence of transcript assembly on the proteogenomics discovery of microproteins. *PLoS One*, 13(3):e0194518, 2018.
- [49] Dharendra Kumar, Amit Kumar Yadav, Xinying Jia, Jason Mulvenna, and Debasis Dash. Integrated transcriptomic-proteomic analysis using a proteogenomic workflow refines rat genome annotation. *Molecular & Cellular Proteomics*, 15(1):329–339, 2016.

- [50] Michal Levin, Marion Scheibe, and Falk Butter. Proteotranscriptomics assisted gene annotation and spatial proteomics of *bombyx mori* bmn4 cell line. *BMC genomics*, 21:1–14, 2020.
- [51] Paolo Cifani, Avantika Dhabaria, Zining Chen, Akihide Yoshimi, Emily Kawaler, Omar Abdel-Wahab, John T Poirier, and Alex Kentsis. Proteomegenerator: a framework for comprehensive proteomics based on de novo transcriptome assembly and high-accuracy peptide mass spectral matching. *Journal of proteome research*, 17(11):3681–3692, 2018.
- [52] Torsten Müller, Etienne Boileau, Sweta Talyan, Dorothea Kehr, Karl Varadi, Martin Busch, Patrick Most, Jeroen Krijgsveld, and Christoph Dieterich. Updated and enhanced pig cardiac transcriptome based on long-read rna sequencing and proteomics. *Journal of molecular and cellular cardiology*, 150:23–31, 2021.
- [53] TS Keshava Prasad, Ajeet Kumar Mohanty, Manish Kumar, Sree-lakshmi K Sreenivasamurthy, Gourav Dey, Raja Sekhar Nirujogi, Sneha M Pinto, Anil K Madugundu, Arun H Patil, Jayshree Advani, et al. Integrating transcriptomic and proteomic data for accurate assembly and annotation of genomes. *Genome research*, 27(1):133–144, 2017.
- [54] Ceereena Ubaida Mohien, David R Colquhoun, Derrick K Mathias, John G Gibbons, Jennifer S Armistead, Maria C Rodriguez, Mario Henry Rodriguez, Nathan J Edwards, Jürgen Hartler, Gerhard G Thallinger, et al. A bioinformatics approach for integrated transcriptomic and proteomic comparative analyses of model and non-sequenced anopheline vectors of human malaria parasites. *Molecular & cellular proteomics*, 12(1):120–131, 2013.
- [55] Vanessa C Evans, Gary Barker, Kate J Heesom, Jun Fan, Conrad Bessant, and David A Matthews. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nature methods*, 9(12):1207–1211, 2012.
- [56] Isabel Desgagné-Penix, Morgan F Khan, David C Schriemer, Dustin Cram, Jacek Nowak, and Peter J Facchini. Integration of deep transcriptome and proteome analyses reveals the components of alkaloid metabolism in opium poppy cell cultures. *BMC plant biology*, 10(1):1–17, 2010.
- [57] Xiaoan Lang, Na Li, Lingfei Li, Shouzhou Zhang, et al. Integrated metabolome and transcriptome analysis uncovers the role of antho-

- cyanin metabolism in *Michelia maudiae*. *International Journal of Genomics*, 2019, 2019.
- [58] Wen-Hsiung Li, Chung-I Wu, and Chi-Cheng Luo. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular biology and evolution*, 2(2):150–174, 1985.
- [59] Jinfeng Liu, Yan Zhang, Xingye Lei, and Zemin Zhang. Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. *Genome biology*, 9:1–17, 2008.
- [60] Fumio Tajima. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3):585–595, 1989.
- [61] Pardis C Sabeti, David E Reich, John M Higgins, Haninah ZP Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, Jill V Platko, Nick J Patterson, Gavin J McDonald, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, 2002.
- [62] Benjamin F Voight, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. A map of recent positive selection in the human genome. *PLoS biology*, 4(3):e72, 2006.
- [63] Justin C Fay and Chung-I Wu. Hitchhiking under positive darwinian selection. *Genetics*, 155(3):1405–1413, 2000.
- [64] Carina F Mugal, Jochen BW Wolf, and Ingemar Kaj. Why time matters: codon evolution and the temporal dynamics of d_n/d_s . *Molecular biology and evolution*, 31(1):212–231, 2014.
- [65] Stanley A Sawyer and Daniel L Hartl. Population genetics of polymorphism and divergence. *Genetics*, 132(4):1161–1176, 1992.
- [66] Justin C Fay, Gerald J Wyckoff, and Chung-I Wu. Positive and negative selection on the human genome. *Genetics*, 158(3):1227–1234, 2001.
- [67] Robert E Hill and Nicholas D Hastie. Accelerated evolution in the reactive centre regions of serine protease inhibitors. *Nature*, 326(6108):96–99, 1987.
- [68] Austin L Hughes and Masatoshi Nei. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335(6186):167–170, 1988.

- [69] Zhongming Zhao, Yun-Xin Fu, David Hewett-Emmett, and Eric Boerwinkle. Investigating single nucleotide polymorphism (snp) density in the human genome and its implications for molecular evolution. *Gene*, 312:207–213, 2003.
- [70] Sergey Kryazhimskiy and Joshua B Plotkin. The population genetics of dn/ds. *PLoS genetics*, 4(12):e1000304, 2008.
- [71] Ziheng Yang. The power of phylogenetic comparison in revealing protein function. *Proceedings of the National Academy of Sciences*, 102(9):3179–3180, 2005.
- [72] Maria Anisimova, Joseph P Bielawski, and Ziheng Yang. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular biology and evolution*, 18(8):1585–1592, 2001.
- [73] Ziheng Yang. Phylogenetic analysis by maximum likelihood (paml), 2000.
- [74] Hidetoshi Shimodaira and Masami Hasegawa. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular biology and evolution*, 16(8):1114, 1999.
- [75] Ziheng Yang. Maximum-likelihood models for combined analyses of multiple sequence data. *Journal of Molecular Evolution*, 42(5):587–596, 1996.
- [76] Wendy SW Wong, Ziheng Yang, Nick Goldman, and Rasmus Nielsen. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*, 168(2):1041–1051, 2004.
- [77] Ziheng Yang and Rasmus Nielsen. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular biology and evolution*, 19(6):908–917, 2002.
- [78] Jianzhi Zhang, Rasmus Nielsen, and Ziheng Yang. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution*, 22(12):2472–2479, 2005.
- [79] Daniel C Jeffares, Bartłomiej Tomiczek, Victor Sojo, and Mario dos Reis. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. In *Parasite genomics protocols*, pages 65–90. Springer, 2015.

- [80] PJ Esteves, J Abrantes, M Carneiro, A Müller, G Thompson, and W Van der Loo. Detection of positive selection in the major capsid protein vp60 of the rabbit haemorrhagic disease virus (rhDV). *Virus research*, 137(2):253–256, 2008.
- [81] Alexandra Pavlova, Han Ming Gan, Yin Peng Lee, CM Austin, Dean M Gilligan, Mark Lintermans, and Paul Sunnucks. Purifying selection and genetic drift shaped pleistocene evolution of the mitochondrial genome in an endangered australian freshwater fish. *Heredity*, 118(5):466–476, 2017.
- [82] Ziheng Yang, Wendy SW Wong, and Rasmus Nielsen. Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular biology and evolution*, 22(4):1107–1118, 2005.
- [83] Ashley Lu and Stéphane Guindon. Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. *Molecular biology and evolution*, 31(2):484–495, 2014.
- [84] Bo Xu and Ziheng Yang. Pamlx: a graphical user interface for paml. *Molecular biology and evolution*, 30(12):2723–2724, 2013.
- [85] Steven N Steinway, Ruth Dannenfelser, Christopher D Laucius, James E Hayes, and Sudhir Nayak. Jcoda: a tool for detecting evolutionary selection. *BMC bioinformatics*, 11(1):1–9, 2010.
- [86] Fei Su, Hong-Yu Ou, Fei Tao, Hongzhi Tang, and Ping Xu. Psp: rapid identification of orthologous coding genes under positive selection across multiple closely related prokaryotic genomes. *BMC genomics*, 14(1):1–10, 2013.
- [87] Joel Buset, Cedric Cabau, Camille Meslin, and Geraldine Pascal. Phyleasprog: a user-oriented web server for wide evolutionary analyses. *Nucleic acids research*, 39(suppl_2):W479–W485, 2011.
- [88] B Korber, Allen G Rodrigo, and Gerald H Learn. Computational analysis of hiv molecular sequences. *HIV signature and sequence variation analysis*, 55, 2000.
- [89] Sergei L Kosakovsky Pond and Simon DW Frost. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, 21(10):2531–2533, 2005.
- [90] Etienne Lord, Mickael Leclercq, Alix Boc, Abdoulaye Banire Diallo, and Vladimir Makarenkov. Armadillo 1.1: an original workflow plat-

- form for designing and conducting phylogenetic analysis and simulations. *PloS one*, 7(1):e29903, 2012.
- [91] Emanuel Maldonado, Kartik Sunagar, Daniela Almeida, Vitor Vasconcelos, and Agostinho Antunes. Impact_s: integrated multiprogram platform to analyze and combine tests of selection. *PloS one*, 9(10):e96243, 2014.
- [92] Sébastien Moretti, Riccardo Murri, Sergio Maffioletti, Arnold Kuzniar, Briséis Castella, Nicolas Salamin, Marc Robinson-Rechavi, and Heinz Stockinger. gcodeml: a grid-enabled tool for detecting positive selection in biological evolution. In *HealthGrid*, pages 59–68, 2012.
- [93] Jorge A Hongo, Giovanni M de Castro, Leandro C Cintra, Adhemar Zerlotini, and Francisco P Lobo. Potion: an end-to-end pipeline for positive darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. *BMC genomics*, 16(1):1–16, 2015.
- [94] Andrew E Webb, Thomas A Walsh, and Mary J O’Connell. Vespa: very large-scale evolutionary and selective pressure analyses. *PeerJ Computer Science*, 3:e118, 2017.
- [95] Larry Wall et al. The perl programming language, 1994.
- [96] R Core Team et al. R: A language and environment for statistical computing. *cran.microsoft.com*, 2013.
- [97] Stephen Richard Bourne. Unix time-sharing system: The unix shell. *The Bell System Technical Journal*, 57(6):1971–1990, 1978.
- [98] Romain A Studer and Marc Robinson-Rechavi. How confident can we be that orthologs are similar, but paralogs differ? *Trends in Genetics*, 25(5):210–216, 2009.
- [99] Victoria Nembaware, Karen Crum, Janet Kelso, and Cathal Seoighe. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. *Genome research*, 12(9):1370–1376, 2002.
- [100] Romain A Studer, Simon Penel, Laurent Duret, and Marc Robinson-Rechavi. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome research*, 18(9):1393–1402, 2008.
- [101] Fyodor A Kondrashov, Igor B Rogozin, Yuri I Wolf, and Eugene V Koonin. Selection in the evolution of gene duplications. *Genome biology*, 3(2):1–9, 2002.

- [102] Gregory Jordan and Nick Goldman. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Molecular biology and evolution*, 29(4):1125–1139, 2012.
- [103] Ziheng Yang and Mario Dos Reis. Statistical properties of the branch-site test of positive selection. *Molecular biology and evolution*, 28(3):1217–1228, 2010.
- [104] Penka Markova-Raina and Dmitri Petrov. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 drosophila genomes. *Genome research*, 21(6):863–874, 2011.
- [105] Mikita Suyama, David Torrents, and Peer Bork. Pal2nal: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research*, 34(suppl_2):W609–W612, 2006.
- [106] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.
- [107] Kazutaka Katoh and Daron M Standley. Mafft: iterative refinement and additional methods. *Multiple sequence alignment methods*, pages 131–146, 2014.
- [108] Kuo-Bin Li. Clustalw-mpi: Clustalw analysis using distributed and parallel computing. *Bioinformatics*, 19(12):1585–1586, 2003.
- [109] William Fletcher and Ziheng Yang. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Molecular biology and evolution*, 27(10):2257–2267, 2010.
- [110] Sabrina Dietz, Miguel Vasconcelos Almeida, Emily Nischwitz, Jan Schreier, Nikenza Viceconte, Albert Fradera-Sola, Christian Renz, Alejandro Ceron-Noriega, Helle D Ulrich, Dennis Kappei, et al. The double-stranded dna-binding proteins tebp-1 and tebp-2 form a telomeric complex with pot-1. *Nature Communications*, 12(1):2668, 2021.
- [111] Marie Delattre and Nathan W Goehring. The first steps in the life of a worm: Themes and variations in asymmetric division in *c. elegans* and other nematodes. In *Current Topics in Developmental Biology*, volume 144, pages 269–308. Elsevier, 2021.
- [112] Danielle R Hamill, Aaron F Severson, J Clayton Carter, and Bruce Bowerman. Centrosome maturation and mitotic spindle assembly in *c. elegans* require spd-5, a protein with multiple coiled-coil domains. *Developmental cell*, 3(5):673–684, 2002.

- [113] Jeffrey B Woodruff, Oliver Wueseke, and Anthony A Hyman. Pericentriolar material structure and dynamics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1650):20130459, 2014.
- [114] Lisa Stenzel, Alina Schreiner, Elisa Zuccoli, Sim Üstüner, Judith Mehler, Esther Zanin, and Tamara Mikeladze-Dvali. Pcmd-1 bridges the centrioles and the pericentriolar material scaffold in *c. elegans*. *Development*, 148(20):dev198416, 2021.
- [115] Morkos A Henen, Walter Myers, Lauren R Schmitt, Kristen J Wade, Alexandra Born, Parker J Nichols, and Beat Vögeli. The disordered spindly c-terminus interacts with rzz subunits rod-1 and zwl-1 in the kinetochore through the same sites in *c. elegans*. *Journal of molecular biology*, 433(4):166812, 2021.
- [116] Richard Bayliss, Teresa Sardon, Isabelle Vernos, and Elena Conti. Structural basis of aurora-a activation by tpx2 at the mitotic spindle. *Molecular cell*, 12(4):851–862, 2003.
- [117] Anne-Lore Schlaitz, Martin Srayko, Alexander Dammermann, Sophie Quintin, Natalie Wielsch, Ian MacLeod, Quentin de Robillard, Andrea Zinke, John R Yates, Thomas Müller-Reichert, et al. The *c. elegans* rsa complex localizes protein phosphatase 2a to centrosomes and regulates mitotic spindle assembly. *Cell*, 128(1):115–127, 2007.
- [118] Philippe Jarne. Mating system, bottlenecks and genetic polymorphism in hermaphroditic animals. *Genetics Research*, 65(3):193–207, 1995.
- [119] Sungsu Park, Hyejin Hwang, Seong-Won Nam, Fernando Martinez, Robert H Austin, and William S Ryu. Enhanced caenorhabditis elegans locomotion in a structured microfluidic environment. *PloS one*, 3(6):e2550, 2008.
- [120] Alex Gomez-Marin, Greg J Stephens, and André EX Brown. Hierarchical compression of caenorhabditis elegans locomotion reveals phenotypic differences in the organization of behaviour. *Journal of The Royal Society Interface*, 13(121):20160466, 2016.

9 Curriculum Vitae

