Ultraconserved elements coupled with machine learning approaches resolve the systematics in model nematode species

Villegas, Laura, worm∼lab, IfZ, University of Cologne
Jimenez, Lucy, worm∼lab, IfZ, University of Cologne
van der Sprong, Joëlle, Ludwigs Maximilian University of Munich
Holovachov, Oleksandr, Swedish Museum of Natural History
Waldvogel, Ann-Marie, Technical University of Munich
Schiffer, Philipp H., worm∼lab, IfZ, University of Cologne

May 6, 2025

Abstract

Nematodes are among the most diverse animal groups, inhabiting nearly all terrestrial and aquatic ecosystems. More than a million species of nematodes are expected to occur on Earth. However, only around 28,000 have been described to date. Nematode phylogenetics remains challenging due to their small size, morphological simplicity, and cryptic diversity. Traditional morphological and molecular approaches, such as single-locus markers (e.g., 18S rRNA, COI), often lack resolution, particularly at shallow evolutionary scales. Moreover, morphology-based classifications are flawed, and phylogenies inferred from single-locus data often fail to resolve deep branching relationships within the nematode phylogenetic tree. These limitations highlight the need for more comprehensive genomic approaches to achieve higher-resolution evolutionary inferences. Here, we design and test the first ultraconserved elements (UCEs) probe set for two nematode families: Panagrolaimidae and Rhabditidae. This approach captures thousands of loci without requiring whole-genome or transcriptome sequencing. Our probe sets targeted 1612 and 100397 UCE loci for Panagrolaimidae and Rhabditidae, respectively. In vivo testing for Panagrolaimidae captured up to 1457 loci, enabling a robust phylogenetic reconstruction. Genera classifications within Panagrolaimidae were congruent with prior phylogenies, except for one strain, which we here redescribe as *Neocephalobus halophilus* BSS8, based on morphological and molecular evidence. We applied machine learning classifiers to determine the minimum loci required for genus-level classification. For the Rhabditidae family, benchmarking of machine learning models revealed that XGBoost provided the highest accuracy for genus-level classification, with 46 loci as the most informative. For Panagrolaimidae, despite the availability of extensive laboratory isolates, genomic resources are limited. We identified 63 loci as the most informative for the classification of this family. In summary, our UCE probe sets provide a scalable and cost-effective tool for enhancing taxonomic resolution and evolutionary inference in nematodes. This approach has the potential to improve biodiversity assessments and deepen our understanding of this ecologically important group, even with shallow sequencing approaches on-site.

Keywords: Nematoda | ultraconserved element | Phylogenomics | Machine learning | Genus Classification

1 Introduction

Nematodes (Phylum Nematoda) are one of the most diverse and abundant animal taxa, occupying nearly all ecosystems and playing crucial ecological roles, such as regulating microbial and fungal communities, facilitating nutrient cycling, and contributing to soil health and decomposition processes as well as being parasites of plants and animals, including humans (M. Blaxter, 2011; van den Hoogen et al., 2020). Nematode taxonomic diversity remains poorly characterized, with recent estimates of the total number of species in the phylum ranging between half a million and up to 10 million species (Hodda, 2022). As only a little over 28000 species are formally described, an immense 'taxonomic gap' not only limits our understanding of biodiversity, but also constrains our ability to explore the evolutionary history and diversification of this

group. Better resolution of nematode phylogenies is furthermore crucial for advancing the understanding of ecological dynamics, e.g., in soils, where nematode diversity can be linked to patterns on the environmental scale (Villegas et al., 2024).

Resolving phylogenetic placement of species is the basis for reconstructing evolutionary relationships, shedding light on lineage diversification, biogeography, and adaptation. Current approaches, such as morphology-based classification, are often limited by cryptic diversity, convergent evolution, and the need for specialized taxonomic expertise. Similarly, traditional molecular markers (e.g., 18S rRNA, COI) may lack sufficient phylogenetic resolution to reliably distinguish closely related species, while some nematode taxa exhibiting particularly high molecular evolution rates, which can in turn produce skewed or incongruent topologies in phylogenetic reconstructions (M. L. Blaxter et al., 1998; Van Megen et al., 2009). This issue highlights the importance of developing new methods for phylogenetically robust species placement, with a focus on the careful selection of informative loci. Several sequencing methods have emerged as a complement or alternative to amplicon sequencing, like whole genome sequencing (WGS), transcriptome sequencing, target sequence capture, and restriction-site associated sequencing (RAD-seq). Each of these methods offers distinct advantages. WGS enables the comprehensive investigation of complex genomic traits and adaptive processes (Lu et al., 2025), allowing the study of coding and non-coding regions and the analysis of structural variation. However, this comes at increased costs, and data analysis can be computationally intensive. RAD-seq provides lower cost alternative for population and species-level studies when compared to whole genome approaches by focusing on a reduced subset of the genome (K. R. Andrews, Good, Miller, Luikart, & Hohenlohe, 2016), however determining orthology relationships remains challenging. Missing data due to mutations at restriction sites, and potential non-independence of adjacent loci due to linkage disequilibrium are some of the potential drawbacks of this approach (Rubin, Ree, & Moreau, 2012).

For targeted sequence capture methods, specific genomic regions are enriched for sequencing. In this way, a large set of orthologous loci can be selected with probes, making the method particularly useful for phylogenetics and evolutionary studies (M. R. Jones & Good, 2015). However, a major drawback is the requirement of prior genomic or transcriptomic data availability for probe design. Consequently, successful implementation depends on the availability of high-quality genomes, which may be limited for non-model organisms. One common family of bait sets targets highly conserved genomic regions, among them ultra-conserved elements (UCEs), which recover sets of loci that are highly conserved and thus can be captured across divergent groups of organisms. The presence of highly conserved regions (cores) that are anchored between high-variability informative sites (flanking regions) makes it a cost-efficient approach that captures both deep phylogenetic relationships and recent divergences, enabling high-resolution phylogenomic reconstructions (Faircloth et al., 2012). UCEs approaches have been successfully implemented in a wide range of lineages(Blaimer et al., 2015,?; Erickson, Pentico, Quattrini, & McFadden, 2020; Gilbert et al., 2015; Quattrini et al., 2017; van der Sprong et al., 2023; Winker, Glenn, & Faircloth, 2018). Beyond evolutionary studies, UCE-based analyses can also provide critical insights into ecologically relevant patterns, such as population structure, connectivity, and demography, which are essential for conservation planning and bio-monitoring efforts (Duckett, Calder, Sullivan, Tank, & Carstens, 2023).

Ultimately, the most suitable method depends on the focus taxonomic group and specific research objectives, as different approaches vary in their effectiveness depending on genome complexity, evolutionary scale, and data requirements. In this study, we designed two ultraconserved elements (UCEs) bait sets for nematodes, particularly focusing on the Rhabiditidae and Panagrolaimidae families. Using Rhabditidae for method benchmarking, we obtained phylogenetic reconstructions based on UCE loci that were congruent to those previously proposed in the literature based on orthologs.

For the Panagrolaimidae family, for which only a few genomic resources are available compared to Rhabditidae, we show that using UCEs and target capture provides a high-resolution phylogenetic reconstruction. Additionally, we developed a machine learning framework to identify the most informative UCEs for the taxonomic assignment of genera, offering a scalable and efficient tool for future ecological and evolutionary studies.

The genomic era has generated high-dimensional datasets that pose significant challenges to traditional taxonomy, particularly in groups with high diversity, like nematodes (M. Blaxter, 2011; Hodda, 2022). To effectively interpret these complex data, it is essential to adopt computational approaches that can uncover intricate patterns within genomic variation. The application of machine learning (ML) techniques offers a

promising way to analyse UCE precense-absence paanalyze for presence-absencehe ability of ML algoritby leveragingo identify informative features and build robust predictive models (Libbrecht & Noble, 2015). Notably, presence–absence data have also proven effective for phylogenetic inference, recovering meaningful evolutionary signal even in complex datasets (Natsidis, Kapli, Schiffer, & Telford, 2021). The necessity to select relevant features within these complex dataset requires efficient feature selection methods, which allow for improved model interpretability and reduced dimensionality without compromising the accuracy (Guyon & Elisseeff, 2003; Saeys, Inza, & Larrañaga, 2007). Therefore, the integration of ML and UCE analysis represents a promising strategy to address taxonomic challenges in nematodes, offering a scalable and objective framework for genus-level classification (Lv, Li, Wang, & Zou, 2023). The use of presence–absence patterns of UCEs in this context is of particular interest, as these binary patterns may carry important phylogenetic signal that could improve genus-level classification and contribute to a more accurate understanding of evolutionary relationships. This approach is especially valuable in groups where traditional taxonomic expertise is limited or insufficient, providing a new way to explore complex phylogenetic structures within nematodes.

2 Materials and methods

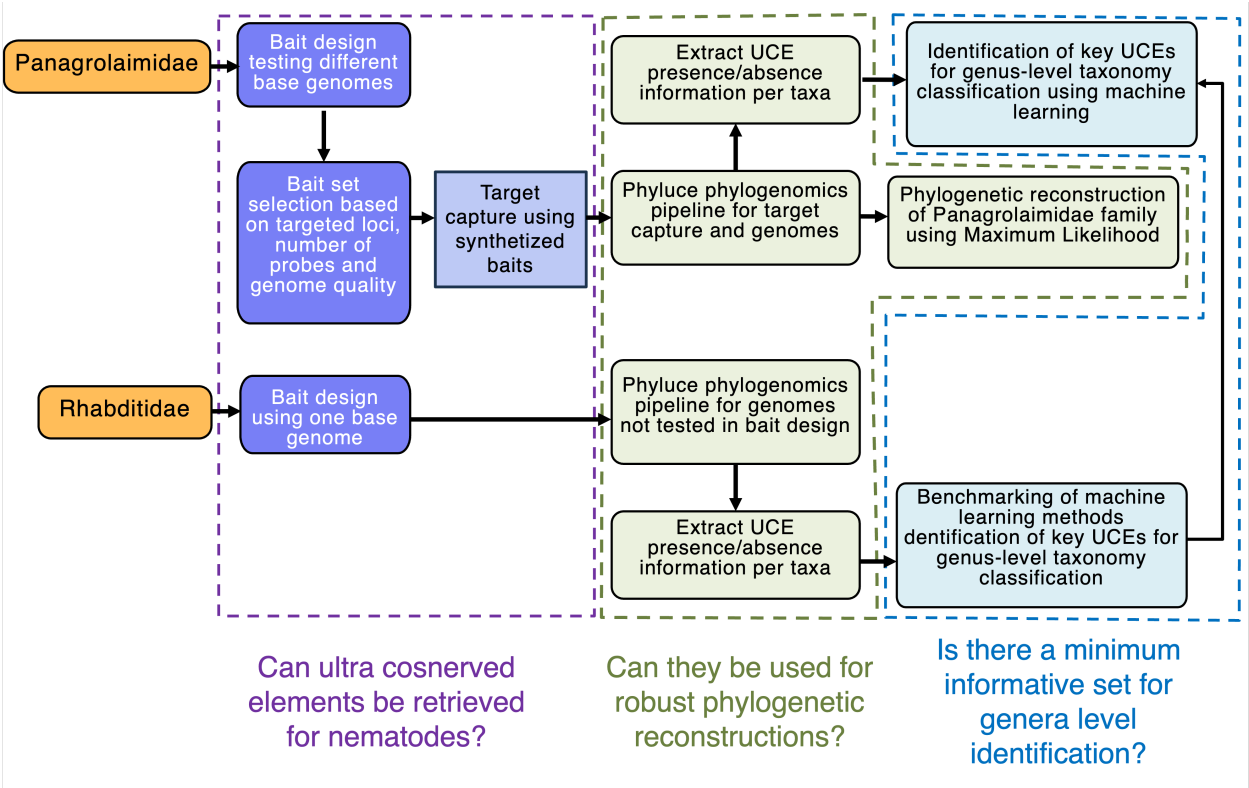


Figure 1: Overview of the workflow implemented in this study. Bait sets were designed and tested for both the Panagrolaimidae and Rhabditidae families. From UCEs harvested in the different families with various approaches, a presence/abscence matrix was obtained to train and test machine learning classifiers to identify the most important UCEs for genera level classification. For Panagrolaimidae, the model was further tested using UCE data extracted from genome skims generated with Nanopore sequencing. Graph generated based on canva template from elversa.

2.1 Phylogenomic analysis of the family Panagrolaimidae: in-silico and in-vivo testing

2.1.1 Bait set design and in-silico testing

Bait set design

Phyluce 1.7.2 was used for the bait set design, the "Tutorial IV: Identifying UCE Loci and Designing Baits To Target Them" was followed (phyluce.readthedocs) (Faircloth, 2015; Faircloth et al., 2012). Eight base genomes were tested for the bait set design: *Panagrolaimus sp.* PS1159 (GenBank accession number: GCA_901765195.1), *Panagrolaimus* sp. ES5, *Panagrolaimus* sp.. PS1579 (GenBank accession number: GCA_901779485.1), *Panagrolaimus kolymaensis* (GenBank accession number:GCA_028622995.1), *Panagrolaimus* sp. LJ2400 (GenBank accession number: GCA_024447215.1), *Panagrolaimus* sp. LJ2406 (GenBank accession number: GCA_024447205.1), *Panagrolaimus* sp. LJ2414 (GenBank accession number: GCA_024447195.1) and *Panagrolaimus superbus* (GenBank accession number: GCA_901766145.1).

Bait sets were individually designed using each of the base genomes and further assessed based on the quality of the base genome, the number of conserved loci consistently found between exemplar taxa, and the final probe count (for further synthesis).

Illumina reads for each of the strains used as base genomes were obtained through the Sequence Read Archive (SRA) with the exception of *Panagrolaimus kolymaensis* which was sequenced using a No,vaSeq illumina platform 2x150bp by NoIlluminaUK) Company Limited. Publicly available re-sequencing data for the other strains tested were obtained through the Sequence Read Archive (SRA) (supplementary table 3).

Reads were mapped against the base genomes using stampy (v 1.0.32) (Lunter & Goodson, 2010) allowing for sequence divergence ≤ 5. Unmapped reads were removed using samtools (v 1.18) (Danecek et al., 2021). Filtered bam files were converted to bed format, and overlapping intervals were merged using bedtools (v 2.31.1) (Quinlan & Hall, 2010) for putatively conserved regions. Repetitive intervals and ambiguous bases were removed using phyluce_probe_strip_masked_loci_from_set, intervals shorter than 80bp and 25% of the regions were masked using the same tool. Sequences with the identified UCEs were extracted using phyluce_probe_get_genome_sequences_from_bed specifying a buffer region of 160 bp. Temporary bait sets targeting those sequences were obtained using phyluce_probe_get_tiled_probes with at tilling density set to three, using the `−masking`, `−remove−gc` flags, potentially problematic baits with over 25% repeat content and GC content outside of 30-70% were also removed. Duplicate baits were screened and removed using phyluce_probe_easy_lastz and phyluce_probe_remove_duplicate_hits_from_probes_using_lastz.

The baits were then aligned against each of the tested base genomes using phyluce_probe_run_multiple_lastzs_sqlite with an identity value of 50%. Baits that matched multiple contigs were removed. Using phyluce_probe_slice_sequence_from_genomes, fasta sequences from each exemplar taxa were extracted buffering each locus to 180 bp. The final bait design was obtained using phyluce_probe_get_tiled_probe_from_multiple_inputs, the baits were designed with a tilling-density of three, baits with 25% masking were removed, and two probes targeting each locus were design. Duplicate baits were removed (with an identity value of 50%) using phyluce_probe_easy_lastz and phyluce_probe_remove_duplicate_hits_from_probes_using_lastz.

The different bait lists were further examined by the bioinformatics team from the company BIOCAT, and the selected bait set was synthesized using myBaits® Custom (1–20000) target capture kit by Daicel Arbor BioSciences (Ann Arbor, MI, USA).

2.1.2 Sampling, DNA extraction and sequencing

Nematodes were harvested from cultures in agar plates by washing off plates with 2mL of nuclease free water, pelleted by centrifuging at 4 degrees for 5 min at 3000g. Nematodes are kept as laboratory cultures at 15°C on low nutrient agar plates inoculated with OP50 *(E. coli)*. Reproductive mode assessment has previously been performed by (Lewis et al., 2009; Villegas et al., 2024) on a set of the strains here tested. DNA extractions for in-vitro testing of the baits were performed using the Quick-DNA Microprep Plus Kit (catalogue number: D4074) manufacturer's protocol for tissue samples was modified by increasing proteinase k volume to 20μL per sample and incubating overnight for 20 hours. Sequencing was performed on a NovaSeq 6000 Illumina sequencer with 2x150bp at the Cologne Center for Genomics (CCG). DNA extractions, library preparation and sequencing were done for 22 strains.

2.1.3 Analysis of sequencing data: in-vivo testing and phylogenomic reconstruction

Sequencing data resulting from target capture was analysed following the "Tutorial I: UCE Phylogenomics" (phyluce.readthedocs) with Phyluce 1.7.3. Sequencing adapters were trimmed using Cutapadt (v 4.9)(M. Martin, 2011)(supplementary table with adapters list). Quality of the trimmed reads was assessed using FastQC (v0.11.2) (S. Andrews, 2010). Assemblies for each strain were conducted using Spades (v 3.14.1) (Prjibelski, Antipov, Meleshko, Lapidus, & Korobeynikov, 2020) incorporated in phyluce (phyluce_assembly_assemblo_spades). Quality of the assemblies was checked using phyluce_assembly_get_fasta_lengths. Genomes available through NCBI for the Panagrolaimidae family were used to obtain a more comprehensive phylogenetic reconstruction of this family and test the robustness of using UCEs as a smaller subset of the genome for phylogenetic reconstructions (supplementary table 4). UCE loci from genomes were harvested following the tutorial "Tutorial III: Harvesting UCE Loci From Genome"(phyluce.readthedocs. Genomes were converted into 2bit format using faToTwoBit (v 445 and 472) (R. M. Kuhn, Haussler, & Kent, 2012). The UCE probe set was aligned to the genomes and the fasta sequence corresponding to the loci were extracted using phyluce_probe_run_multiple_lastzs_sqlite and phyluce_probe_slice_sequence_from_genomes. The extracted fasta sequences were then processed along with the UCE loci obtained from the spades assemblies for the sequenced target capture data.

For known triploid taxa, we ran dedupe.sh from BBMap (v 39.08) (Bushnell, 2014) to avoid having duplicated UCE loci as a result of an uncollapsed genome assembly. We then matched the assembled contigs and available genomes of each strain to the UCE baits using phyluce_assembly_match_contigs_to_probes. The UCE loci were then extracted using phyluce_assembly_get_fastas_from_match_counts. In the case of *Panagrolaimus kolymaensis*, the contigs were fragmented into 2000 bp long and the assembly file was split into 3 different files using a round-robin method since harvesting UCEs from this genome yield high number of duplicated UCEs due to the triploid nature and the deduping method was not useful.

The UCE loci, both harvested from genomes and from target capture data were aligned and edge trimmed using phyluce_align_seqcap_align. The alignments were then cleaned using phyluce_align_remove_locus_name- _from_files. Subsequently, a data matrix of 60% occupancy was created using phyluce_align_get_only_loci_with_min_taxa (each UCE that was present in at least 60% of the taxa analyzed). The alignments were concatenated using phyluce_align_concatenate_alignments. The resulting alignment was used as input for IQ-tree (v 2.3.6) (Minh et al., 2020), a maximum likelihood inference was made using ultra fast bootstrap (1000 bootstrap replicates) with the Model Finder Plus feature for automatic selection of the best substitution model for the dataset. The resulting phylogenetic tree was visualized using TreeViewer (v 2.2.0) (Bianchini & Sánchez-Baracaldo, 2024) using the *Further transformation* module to root the tree and change taxa name from accession number and sequencing code to strain or species name, and the *Plot actions* module to display the legend with bootstrap values and add the scale bar.

2.1.4 Species reassessment based on morphological and molecular evidence

For the strain previously identified as *Panagrolaimus detritophagus* BSS8, morphometric measurements were taken and compared with those of closely related species to accurately determine its taxonomic identity. Measures included, among others, body length, vulva location, lip region width, stoma length, corpus length, pharyngeal region length and tail length in both females and males.

For light microscopy, specimens were relaxed using heat, fixed in a cold 4% formaldehyde solution, gradually transferred to pure glycerine using a slow evaporation method, and mounted on permanent slides in glycerine, with paraffin wax used to support the coverslip. Nine male and nine female individuals isolated from current cultures were compared to closely related species: *Neocephalobus halophilus* Paetzold, 1958 (Paetzold, 1958), *Neocephalobus aberrans* Steiner, 1929 (Steiner, 1929a) and *Panagrolaimus orthomici* Korentchenko, 1992 (Korentchenko, 1992) as well as to the description of the original wild population of the same "Panagrolaimus population II" (Bostrom, 1988)

2.2 Ultraconserved elements for accurate species delimitation: test case using Caenorhabditis

Bait-set design and in-silico testing

Phyluce 1.7.3 was used for the bait set design, the "Tutorial IV: Identifying following UCE "ci and Designing Baits To Target Them" was followed (https://phyocs.io/en/latest/tutorials/tutorial-4.htmlphyluce.readthedocs). Only one base genome tested in this case, this correwas sponded to *Cacorrespondingans* (Bristol N2) (GenBank accession number: GCA_000002985_3). The probe design was performed in the same way as for Panagrolaimidae with the exception of using artificially generated reads in this case using illumina_art as specified in the phyluce tutorials instead of real sequencing data. For the generation of this family's bait set the following *Caenorha,bditis* taxa were used to identify putatively conserved regions: *Caenorhabditis brenneri* (GenBank accession number: GCA_964036135.1), *Caenorhabditis briggsae*(GenBank accession number: GCA_021491975.1), *Caenorhabditis elegans* (Bristol N2) (GenBank accession number: GCA_000002985.3), *Caenorhabditis nigoni* (GenBank accession number: GCA_002742825.1), *Caenorhabditis remanei* (GenBank accession number: GCA_001643735.4 and *Caenorhabditis tropicalis* (GenBank accession number: GCA_016735795.1).

The master bait set list was then used to harvest UCEs from 199 genomes available for Rhabditidae through NCBI for the genera *Oscheius*, *Diploscapter*, *Auanema*, and *Caenorhabditis* (supplementary table 10).

Phylogenetic reconstruction of Caenorhabditis

To test for the accuracy of phylogeneticconstructions within Rhabditidae using ultra-conserved elements compared to previously reported methods (e.g. orthologuos genes), a phylogeny,yorthologoushabditis species was reconstructed. A data matrix of 75% occupancy was created using phyluce_align_get_only_loci_with_min_taxa (each UCE that was present in at least 75% of the taxa analyzed). The alignments were concatenated using phyluce_align_concatenate_alignments. The resulting alignment was used as input for IQ-tree (v 2.3.6), a maximum liIQ-TREEd v.nferenc. Aas made using ultra fast bootstrap (1000 oultrafast bootstrapping) the Model Finder Plus feature for Model Finder selection of the best substitution model for the dataset. The resulting phylogenetic tree was compared to (Stevens et al., 2019) phylogenetic reconstruction through a tanglegram obtained using ape (v 5.8-1) (Paradis & Schliep, 2019) and phytools (v 2.4-4) (Revell, 2024).

2.3 Classification model with Ultraconserved elements (UCEs)

2.3.1 UCE data preparation and preprocessing

To analyze the distribution of UCEs, a multi-step workflow was implemented. First, UCE identifiers were extracted from species-specific fasta files for nematodes within the families Panagrolaimidae and Rhabditidae. These identifiers were then used to construct presence-absence matrices, capturing UCE occurrences per strain. It is essential to note that, although we refer to species throughout this section, the dataset comprises strains, with some species represented by multiple strains. Once the presence-absence matrix was constructed, an exploratory data analysis (EDA) was conducted to examine the distribution of UCEs across the samples.

To facilitate subsequent machine learning analyses, species were assigned to their respective genera. This taxonomic assignment was performed using accession numbers retrieved from the European Nucleotide Archive (ENA). The final output consists of genus-level UCE presence-absence matrices, which are the foundation for predictive modeling. For a detailed description of the extraction steps, matrix construction, and taxonomic assignment, refer to Supplementary Information (A.3), where we provide full methodological details, dataset summaries, and the associated scripts.

The data sets used for training and evaluation were derived from genus-level UCE presence-absence matrices, where genera with insufficient representation (frequency ≤ 1) and rows with missing values were removed. After filtering, the Rhabditidae dataset consisted of 197 samples, 8336 features (UCEs), and four genera: *Oscheius*, *Diploscapter*, *Auanema*, and *Caenorhabditis*. For Panagrolaimidae, 1595 predictors (UCEs) were analyzed. Following the removal of genera with insufficient representation and rows with missing values, the Panagrolaimidae dataset (including the outgroup from Cephalobidae - *Acrobeloides*) contained 51 samples and five genera: *Acrobeloides*, *Halicephalobus*, *Panagrellus*, *Panagrolaimus*, and *Propanagrolaimus*. In both datasets, the UCE columns were used as features, and the genus column was converted to a factor for classification, see table 1.

Table 1: Summary of UCE datasets for machine learning.

Family	Genera		Features	Samples
Rhabditidae	Oscheius	49	8336	197
	Diploscapter	3		
	Auanema	8		
	Caenorhabditis	139		
Panagrolaimidae	Acrobeloides	6	1595	51
	Halicephalobus	4		
	Panagrellus	3		
	Panagrolaimus	35		
	Propanagrolaimus	2		

2.3.2 Model training, evaluation, and feature selection

The Rhabditidae dataset, which contained the largest number of features and entries, was used to evaluate four machine learning models in R: Random Forest (RF), Logistic Regression (LR), k-Nearest Neighbors (k-NN), and Extreme Gradient Boosting (XGBoost). A stratified partitioning strategy, implemented using the `caret` package (M. Kuhn, 2008), was employed to divide the data into training (65%) and testing (35%) subsets, preserving the original distribution of genera classes in both sets. Class weights were calculated and incorporated into the Random Forest (RF) and XGBoost models to address class imbalance. Model training utilized five-fold cross-validation, with performance metrics calculated using the `multiClassSummary` function from the `caret` package. Random Forest was trained by optimizing the `mtry` parameter with class weights applied, using the `randomForest` package (Liaw & Wiener, 2002). Logistic Regression with PCA involved dimensionality reduction through PCA, followed by feature scaling and regularization via the `glmnet` package (Friedman, Hastie, & Tibshirani, 2010). k-Nearest Neighbors was implemented using default parameters, using the `class` package (Venables & Ripley, 2002). Extreme Gradient Boosting was trained with class weights incorporated into the learning process, using the `xgboost` package (Chen & Guestrin, 2016). Performance metrics, including overall accuracy, kappa statistics, sensitivity, and specificity, were used to compare model performance. A confusion matrix provided detailed insight into classification accuracy for each genus. To identify the most relevant features for nematode genera classification, we employed a feature selection approach based on the importance scores generated by the XGBoost model using the `xgboost` package (Chen & Guestrin, 2016). Specifically, we retained only those features that exhibited a non-zero "Overall" importance score, as determined by the `varImp` function in the `caret` package (M. Kuhn, 2008).

2.3.3 Optimized model application to Panagrolaimidae data

To evaluate the performance of the feature-selected XGBoost model in classifying nematode genera within the Panagrolaimidae family, we applied it to the corresponding UCE presence-absence dataset. The "Genera" column was converted to a factor, and genera represented by single occurrences were removed. Missing values were handled using `na.omit()`. The resulting dataset, which included the following sample sizes per genus: *Acrobeloides* (6), *Halicephalobus* (4), *Panagrellus* (3), *Panagrolaimus* (35), and *Propanagrolaimus* (2), exhibited class imbalance, (see Table 1). To address this, a stratified split of 65% training and 35% testing was performed using `createDataPartition` from the `caret` package (M. Kuhn, 2008), with a fixed seed and manual adjustment to ensure proportional representation. Class weights, inversely proportional to genus frequency, were applied. The XGBoost model was trained using the `train` function from `caret` with the `xgbTree` method, employing 5-fold cross-validation (`cv`, `number = 5`) and `multiClassSummary` for performance evaluation. The Area Under the Curve (AUC) was used as the primary performance metric (`metric = "AUC"`). Subsequently, the model was retrained using only the most important features, identified by an overall importance value greater than 0, to reduce dataset dimensionality. The test set AUC was then calculated using `multiclass.roc` from the `pROC` package (Robin et al., 2011). A confusion matrix was generated to visualize the model's classification performance.

3 Results

3.1 Bait design

The bait set design for Panagrolaimidae was conducted by testing eight distinct base genomes and evaluating the number of conserved loci that were shared between 1 and 9 taxa. The number of loci targeted across different bait sets varied from 0 to 12740. To ensure a balance between taxonomic representation and the number of loci targeted, only bait sets that included more than 1000 loci shared among six or seven taxa were selected for further analysis. In these cases, the number of targeted loci ranged from 1096 to 2080, with the corresponding number of probes ranging from 14018 to 24217 prior to duplicate removal. After duplicates were removed, the final probe count ranged from 13565 to 24006. The Panagrolaimidae bait set was tested on the Rhabditidae family; however, no significant amount of loci could be retrieved. Therefore, a family-specific bait set was designed for Rhabditidae.

The selection of the preferred bait set was based on three criteria: the quality of the base genome, the total number of loci targeted, and the total number of probes. The preferred bait set for synthesis ultimately targeted 1612 loci using 18789 probes.

The bait set designed for Rhabditidae was designed using the *Caenorhabditis elegans* Bristol N2 genome as a base. For the test bait set, six *Caenorhabditis* species were used. The preferred bait set in this case targets 10397 loci shared among the six datasets used. The number of probes prior to duplicate removal was 124612, while the final probe count was 121966.

3.2 In-vivo and in-silico testing Panagrolaimidae

SPAdes assemblies of target capture data ranged between 42479 and 287752 contigs (mean 88628.05 ±56482,50529) for Panagrolaimidae exemplars. The two assemblies from the outgroups Cephalobidae were more fragmented than in the Panagrolaimidae exemplars (1121821 and 1237255 contigs). In general, the mean length of the contigs ranged between 215.19 and 1639.88 base pairs (mean 531.86 ± 269.55).

In total, 51 data sets were analyzed and used to reconstruct the phylogeny of the Panagrolaimidae family, including data sets from the outgroup family Cephalobidae (*Acrobeloides*). In total, 1,572 loci with 694,995 informative sites were used, a mean of 442.11 sites per locus, a 95% confidence interval of 8.11, a minimum of 0 sites and a maximum of 1593 sites, with 396 alignments out of 1572 containing more than 0.6 proportion of taxa (n = 30).

In the Panagrolaimidae analysis, a total of 49 to 84 ultraconserved elements (UCEs) were retrieved from the outgroup and utilized for phylogenomic analysis, regardless of the data source (target capture or available genome assemblies). Within the genus *Panagrolaimus*, the number of UCEs retrieved and incorporated into the phylogenomic analysis ranged from 479 to 1457. Specifically, in the target capture dataset, 66 to 80 UCEs were identified in the outgroup, while 277 to 1200 UCEs were recovered for *Panagrolaimus*. There were 84 UCEs commonly found between all genera analyzed, and 500 UCEs uniquely found in Panagrolaimus.

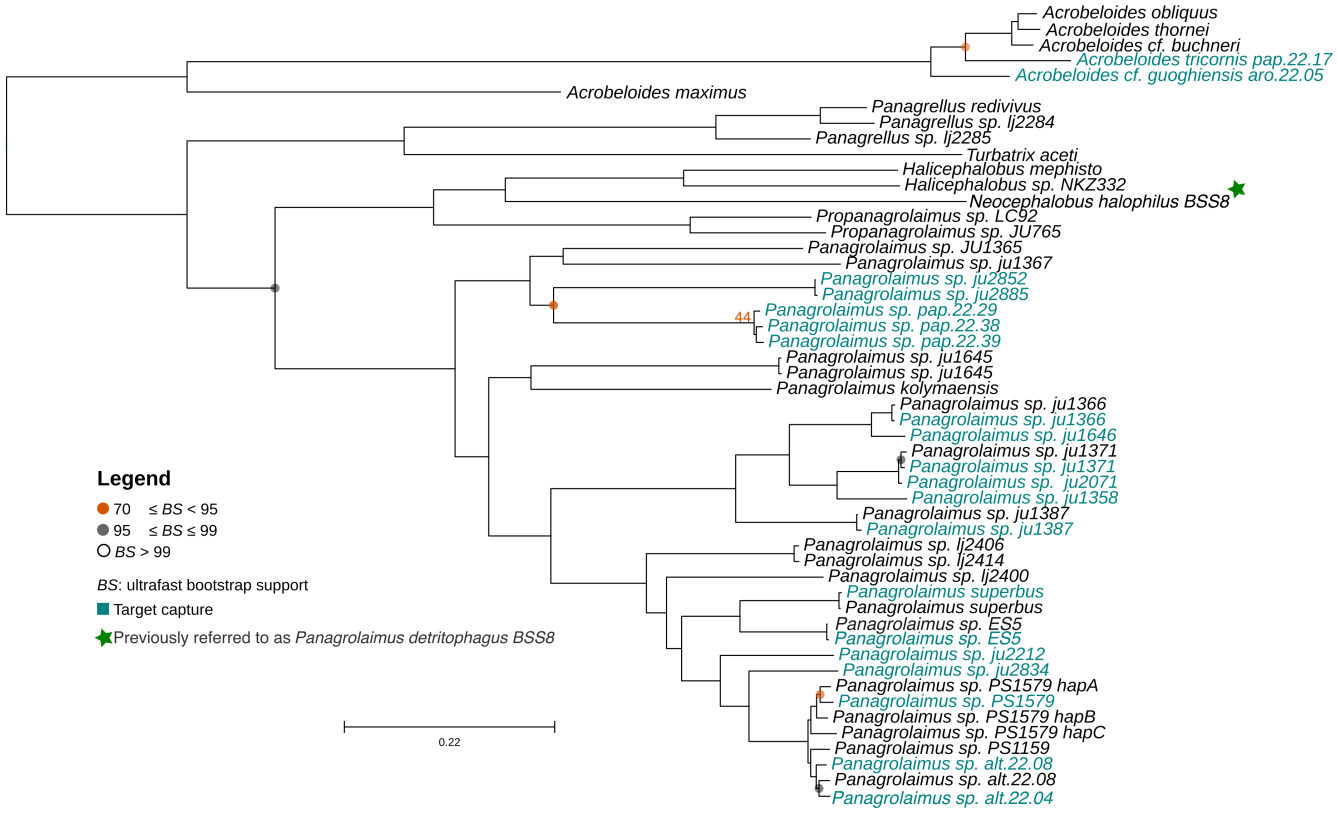


Figure 2: Phylogenetic reconstruction of Panagrolaimidae (*Acrobeloides* outgroup) using maximum likelihood, 51 taxa are included in the analysis. Ultraconserved elements (UCEs) from the target capture approach are highlighted in aquamarine, while UCE data harvested from available genomes are highlighted in black. The phylogeny is based on 396 alignments. Only bootstrap values below 99 are shown in gray and orange.

3.3 Morphological and molecular evidence for species reassessment

Observations of adult individuals, including both females and males of a recent culture of the strain previously referred to as *Panagrolaimus detritophagus* BSS8, revealed that the nematode originally identified as *Panagrolaimus* "population II" in Boström (1988) and subsequently as *Panagrolaimus detritophagus* BSS8 and *Panagrolaimus* BSS8 is in fact *Neocephalobus halophilus* Paetzold, 1958. The main diagnostic feature of *Neocephalobus*, comparing to *Panagrolaimus*, is the presence of distinct papilliform precloacal sensillum in males and its location some distance anterior from the cloacal opening. Additional distinct morphometric features include a longer tail in both males and females in *Neocephalobus* when compared to *Panagrolaimus*, greater pharyngeal region length in *Neocephalobus* when compared to *Panagrolaimus*, and greater body length - body width ratio at mid-body in *Neocephalobus* when compared to *Panagrolaimus* (Supplementary tables 5 and 6 and supplementary text **Morphological description of *Neocephalobus halophilus* BSS8 strain, previously referred to as *Panagrolaimus detritophagus* BSS8**). Furthermore, its placement of the phylogenetic reconstruction based on UCEs places it as a close relative to *Halicephalobus* species and *Propanagrolaimus* rather than *Panagrolaimus* with a bootstrap support value of 100 (figure 2).

3.4 In-silico testing Rhabditidae

The bait set designed with *Caenorhabditis* genomes was then tested on several genera within the Rhabditidae family (*Oscheius*, *Diploscapter*, *Auanema*, and *Caenorhabditis*). The number of UCE loci harvested from the genomes ranged from 1 (*Oscheius myriophilus*) to 5700 (*Caenorhabditis elegans*). Only 50 UCEs were shared between all genera of the family, with most UCEs being uniquely found in *Caenorhabditis*.

In total, 27 data sets were analyzed and used to reconstruct the phylogeny of the genus *Caenorhabditis*. In total, 945 loci with 437091 informative sites were used, a mean of 462.53 sites per locus, a 95% confidence interval of 10.03, a minimum of 33 sites and a maximum of 829 sites, with 945 alignments containing more than 0.75 proportion of taxa.

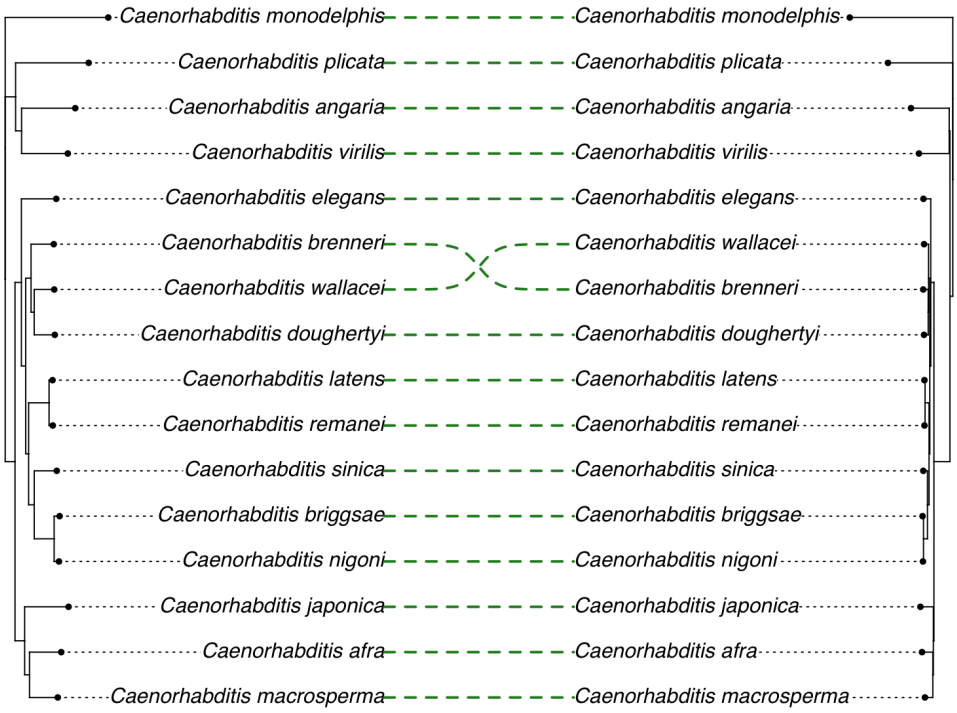


Figure 3: Tanglegram for the genus *Caenorhabditis* comparing UCE derived reconstruction (left) obtained in this work and gene orthology based reconstruction (right) from (Stevens et al., 2019). The alternative topology for *C. wallacei* and *C: brenneri* is also proposed in (Stevens et al., 2019) depending on the method used.

3.5 Classification model with ultraconserved Elements (UCEs)

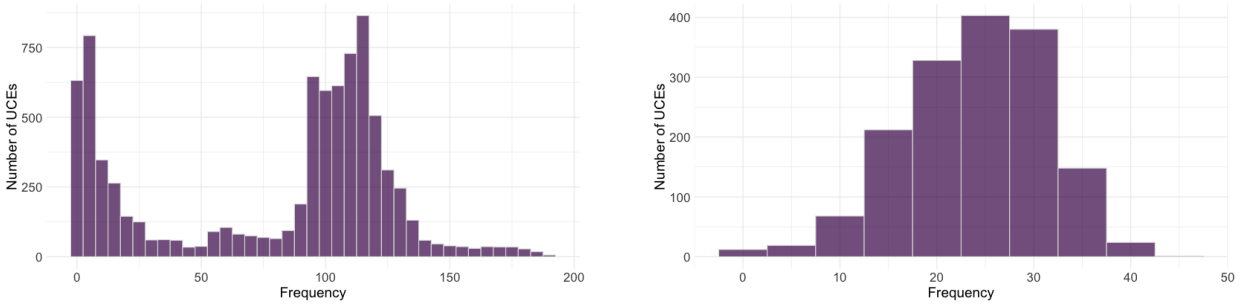
3.5.1 Distribution and characteristics of UCEs in Rhabditidae and Panagrolaimidae

We analyzed the total number of UCEs per strain within the families Rhabditidae and Panagrolaimidae to assess their distribution across strains. The results, visualized in the Supplemental Material (Figure 12 and 11), indicate substantial variation in UCE counts between strains within each family. Although most strains exhibit comparable UCE numbers, some show significantly lower counts. Specifically, in Rhabditidae, UCE counts range from 26 to 11400, with a median of 7762, while in Panagrolaimidae, they range from 15 to 1457, with a median of 767.5. This variability may reflect differences in genome characteristics, such as assembly completeness or lineage-specific variation in the retention of UCEs. This variability may also stem from the bait set design, which was based only on a few taxa, for instance, for Rhabditidae, the bait set was obtained only by analyzing the genus *Caenorhabditis*, yet later applied to a broader range of lineages within Rhabditidae, including genera such as *Diploscapter* that were not represented in the initial design.

To ensure data quality, we applied an outlier detection approach based on UCE counts. Instead of relying solely on interquartile range (IQR) methods, which can be sensitive to skewed distributions, we opted for a more robust approach by removing the bottom 1% of strains with the lowest UCE counts. This method is

particularly useful for genomic datasets where sequencing depth and assembly completeness vary, ensuring that only strains with sufficiently high UCE representation are retained for subsequent analyses. After this procedure, three strains were identified as outliers and removed, two from Rhabditidae (GCA_002207785.1 and GCA_964036155.1 - *Diploscapter coronatus*) and one from Panagrolaimidae (GCA_028622995.1 - *Panagrolaimus kolymaensis* prior to splitting contigs).

After filtering, we analyzed the frequency of UCE occurrences across the remaining strains. The frequency distributions of UCE for both families are presented (Fig 4), showing the number of UCEs shared across different proportions of strains. In Panagrolaimidae, the distribution is unimodal, with most UCEs appearing in an intermediate range of strains. In contrast, the Rhabditidae distribution exhibits bimodality, indicating two groups of UCEs with different levels of conservation. The low-frequency group (0-50) may represent UCEs that are present in some species but absent in others, reflecting potential lineage-specific losses or variation in presence or absence. Conversely, the high-frequency group (100-150) comprises UCEs that are widely shared across Rhabditidae, suggesting they may be under stronger evolutionary constraints, potentially due to their regulatory or structural functions.



(a) Rhabditidae

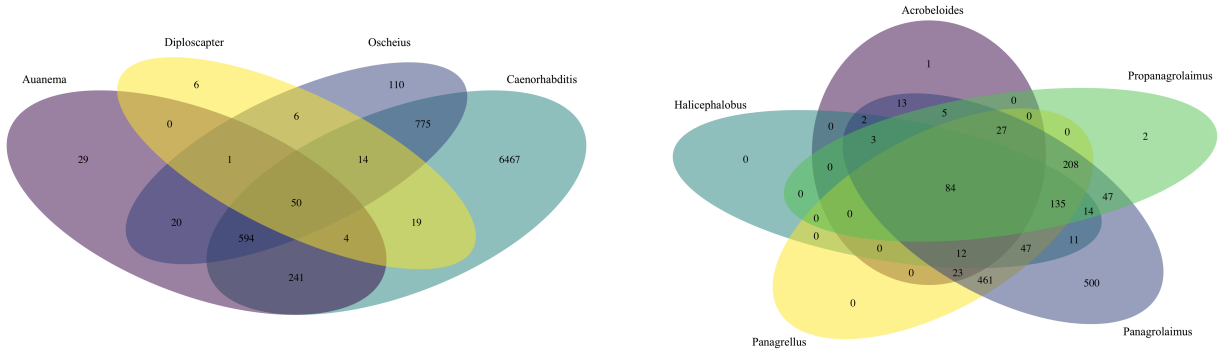
(b) Panagrolaimidae

Figure 4: Divergent patterns of UCE conservation across strains in Rhabditidae and Panagrolaimidae. The x-axis represents the percentage of strains in which a given UCE is present, while the y-axis indicates the number of UCEs observed at each frequency level. The unimodal pattern in Panagrolaimidae contrasts with the bimodal distribution in Rhabditidae, suggesting differences in evolutionary constraints and genomic conservation across the two families.

We analyzed the number of UCEs shared across genera within each family. In Rhabditidae, a total of 50 UCEs are shared across the genera *Auanema*, *Caenorhabditis*, *Diploscapter*, and *Oscheius* (Fig 5a). In Panagrolaimidae, 84 UCEs are shared across the genera *Panagrolaimus*, *Halicephalobus*, *Panagrellus*, *Propanagrolaimus* and the outgroup genus *Acrobeloides* (Fig 5b). The corresponding Venn diagrams summarizing these overlaps are presented in Fig 5, and the full list of shared UCEs is available in Supplementary Table 11

3.5.2 Benchmark of machine learning models on Rhabditidae data

The performance evaluation of the four machine learning models: Random Forest (RF), Logistic Regression (LR) with data transformed with PCA, k-Nearest Neighbors (k-NN), and Extreme Gradient Boosting (XGBoost) on the Rhabditidae dataset of UCEs, along with their respective performance metrics, is shown in Table 2. A comparative analysis revealed differences in classification performance among the models. XGBoost achieved the highest AUC score of 0.9997, demonstrating superior predictive power. It correctly classified all *Auanema* and *Diploscapter* samples and misclassified only one *Oscheius* sample as *Caenorhabditis*. Its strong performance is attributed to its gradient boosting framework and the integration of class weights, which effectively handled class imbalance. Random Forest performed well, achieving an AUC of 0.9939, but struggled with *Oscheius* and *Diploscapter*. Specifically, it misclassified nine out of sixteen *Oscheius* samples as *Diploscapter*, leading to reduced accuracy in predicting minority classes. While class weighting helped mitigate imbalance, it was not as effective as XGBoost's approach. k-Nearest Neighbors (k-NN) obtained an AUC of 0.9858 and showed difficulty in classifying *Oscheius* and *Diploscapter*. Two *Caenorhabditis* samples were misclassified as *Oscheius*, and one *Oscheius* sample was wrongly predicted as *Diploscapter*, indicating



(a) Rhabditidae

(b) Panagrolaimidae

Figure 5: Venn diagrams for shared UCEs across genera in Rhabditidae and Panagrolaimidae. The diagrams illustrate the number of UCEs shared across the genera within each family.

sensitivity to the choice of `k` and the distance metric in high-dimensional data. Logistic Regression with PCA had the lowest performance, with an AUC of 0.8290. While PCA successfully reduced dimensionality and improved computational efficiency, the model struggled with minority class classification. It failed to predict *Auanema* and *Diploscapter* entirely and misclassified multiple *Oscheius* samples, demonstrating limitations despite regularization techniques to prevent overfitting. Overall, XGBoost emerged as the most effective model, demonstrating the highest classification accuracy and robustness across all classes. Its ability to handle imbalanced datasets, coupled with its efficiency and strong generalization capacity, makes it particularly suited for classifying genera within Rhabditidae. The complete confusion matrices and class-specific performance metrics for each model are provided in the Supplementary Materials A.4.

Table 2: Performance metrics for machine learning models on the Rhabditidae family.

Model	**AUC**	**Accuracy (%)**	**Kappa**	**Sensitivity (%)**	**Specificity (%)**	**Best hyperparameter**
RF	0.9939	85.07	0.6819	85.42	96.21	mtry = 3
LR with PCA	0.8290	94.03	0.8599	74.46	98.55	PCA-transformed
k-NN	0.9858	94.03	0.8652	73.96	97.68	k = 5
XGBoost	0.9997	98.51	0.9657	99.48	99.52	Depth = 6, Learning rate = 0.3

Identification of important features UCEs

Following the application of the non-zero importance score feature selection method on the XGBoost model, a total of 46 features were identified as relevant for nematode genera classification on the Rhabditidae family. These features, along with their corresponding "Overall" importance scores, are detailed in the Supplementary Material Table 12. To provide a concise overview of the feature importance, the top 20 features ranked by their "Overall" importance scores are presented in Supplementary Figure 15. The feature `uce.15118` exhibited the highest importance score (100.00), indicating its central contribution to the model's predictive accuracy. Notably, the top six features (`uce.15118`, `uce.1361`, `uce.2297`, `uce.6206`, `uce.17604`, `uce.16976`) demonstrated significantly higher importance scores compared to the remaining features, suggesting their critical role in discriminating between nematode genera for the familie Rhabditidae. Furthermore, six features were also found to be crucial in the cross-validation with varying features, where the model's AUC significantly improved when these features were included.

3.5.3 Classification performance on Panagrolaimidae UCE data

The optimized XGBoost model worked well on the Panagrolaimidae UCE dataset. To rectify the class imbalance using stratified sampling and weighted training, the model was first trained on all UCE features. Next, the training was done using just the top 63 most important UCEs (i.e., those with an overall importance value >0, see Supplemental Material Table 13). The confusion matrices and classification metrics (Supplementary material A.5) showed that the full model achieved an accuracy of 82.35%, while the model

using the reduced set of features achieved 100% accuracy, correctly classifying all species. Based on these results, it suggests that feature selection improved the model's performance while also reducing dimensionality. Detailed statistics and metrics per class are provided in the supplementary material (Section A.5).

4 Discussion

4.1 Ultraconserved elements can be identified and retrieved for nematodes and can be used for robust phylogenetic reconstructions

We designed two bait sets to capture ultraconserved elements within the phylum Nematoda, with a focus on the families Rhabditidae and Panagrolaimidae. We then identified the most informative UCEs using machine learning classifiers to further aid in nematode classification, utilizing genetic information at high resolution. We demonstrate that using the target-captured loci, a well-resolved phylogeny can be obtained for Panagrolamidae, placing correctly described taxa within the family and showing high support values for most nodes, providing insights into the phylogenetic relationships among taxa that previously had no genetic data available.

Our phylogenetic analysis is consistent with previous studies based on both molecular markers and whole-genome data (Abolafia & Vecchi, 2021; Qing et al., 2024; Shokoohi & Masoko, 2024), with the exception of *Panagrolaimus* BSS8, which we now confirm with both molecular and morphological evidence, that it is in fact *Necephalobus halophilus*, being closely related to the genus *Halicephalobus* in our phylogenetic reconstruction. The nematode genus *Panagrolaimus* appears to be more derived within the Panagrolaimidae family, and it is resolved as the closest relative to the genus *Propanagrolaimus*. Both genera form a monophyletic clade that is closely related to *Halicephalobus*, which in turn is sister to *Panagrellus*. All of these genera are distinct from the outgroup, *Acrobeloides*, which is placed outside of the Panagrolaimidae clade.

We also find asexual Panagrolaimus clustering together (*Panagrolaimus sp.* PS1579, *Panagrolaimus sp.* PS1159 Lewis et al. (2009) and the newly isolated *Panagrolaimus sp.* ALT.22.08) (Villegas et al., 2024)) and separately from sexually reproducing ones (*Panagrolaimus einhardii* and *Panagrolaimus superbus* Lewis et al. (2009)). Furthermore, the asexual nematode *Panagrolaimus kolymanesis*, previously placed as an outgroup to all other *Panagrolaimus* nematodes Shatilovich et al. (2023), is here shown to cluster with *Panagrolaimus JU1645* and does not sit on a branch of its own. We find with high support that sexually reproducing strains from the Atacama Desert Villegas et al. (2024) cluster together and are more closely related to other basal *Panagrolaimus* strains as *Panagrolaimus sp.* JU1367 and *Panagrolaimus sp.* JU1365.

In the Rhabditidae family, we see that our phylogenetic reconstruction using UCEs is congruent when compared to previous work (Stevens et al., 2019), showing a clear distinction between the *Elegans* and *Japonica* groups and both forming the *Elegans* supergroup with *Caenorhabditis monodelphis* as an outgroup. While C. brenneri and C. wallacei do not exhibit a direct one-to-one correspondence between their phylogenies, this alternative hypothesis remains consistent with the phylogenetic trees reported in (Stevens et al., 2019), which were generated using different methods.

While using UCEs in both families is a promising approach for phylogenetic reconstructions and further analysis, it is important to note that these "ultraconserved" elements often exhibit conservation only within relatively narrow taxonomic scales (e.g., within families rather than across orders). This pattern suggests a higher rate of sequence divergence across broader nematode lineages, consistent with the rapid molecular evolution observed in the phylum.

4.2 *Panagrolaimus detritophagus* BSS8 is in fact *Neocephalobus halophilus* BSS8 Paetzold, 1958

The *Panagrolaimus* BSS8 strain is one of the cultures of nematodes established by Dr. Björn Sohlenius from samples from the Surtsey Island (Lewis et al., 2009). There were three different populations of Panagrolaimus collected on Surtsey, established in cultures and described by (Bostrom, 1988), of which the *Panagrolaimus* "population II" matches the morphology of BSS8 strain, in both morphology and in measurements, but especially in the presence of distinct midventral precloacal papilliform sensillum, located at a distance from cloacal opening, at level with spicule manubrium. The population was never identified to species level by

(Bostrom, 1988). Still, it was considered to be most similar to *Panagrolaimus chalcographi* by using the key from (Andrássy, 1984) or *Panagrolaimus detritophagus* by using the (Williams, 1986) grouping. The conclusion of the manuscript was that the population II from Surtsey "appears intermediate between *Panagrolaimus detritophagus* and *Panagrolaimus chalcographi*" (Boström, 1988) but no conclusive decision was made. In subsequent publications, this culture was named (without any obvious justification) as *Panagrolaimus detritophagus* BSS8 (Felix et al., 2000; Lewis et al., 2009; Schiffer et al., 2014; Shannon, Browne, Boyd, Fitzpatrick, & Burnell, 2005). Here we provide additional morphological evidence to confirm the taxonomic identity of the BSS8 strain.

Presence of distinct midventral precloacal papilliform sensillum located at a distance from cloacal opening (at the level of spicule manubrium) is one distinct feature of this population, separating it from most known species currently classified in the genus *Panagrolaimus* and all currently accepted genera of Panagrolaimidae. Only the following three species in Panagrolaimidae are morphologically similar to the BSS8 strain:

Neocephalobus aberrans Steiner, 1929, originally described as Cephalobus (Neocephalobus) aberrans by (Steiner, 1929b)) and later transferred to the genus *Panagrolaimus* (= *Panagrolaimus aberrans* (Steiner, 1929b) Godey, 1963) was found in the feces of a guinea pig, kept at the University of California. No measurements were included in the original description, which is rather incomplete in general. However, *Neocephalobus aberrans* can be easily distinguished from the BSS8 in having narrow long stoma with minute tooth, whereas stoma in BSS8 is broad and the tooth is distinct. Moreover, excretory pore in *N. aberrans* is surrounded by a cuticularised ring (absent in BSS8) and shorter tail in both females and in males.

Neocephalobus halophilus Paetzold, 1958, also later transferred to the genus *Panagrolaimus* and renamed to *Panagrolaimus paetzoldi* Goodey, 1963 (homonym of *Panagrolaimus halophilus* Meyl, 1954), was collected in salt marshes in central Germany. Morphologically, it matches the BSS8 strain in all parameters, including the shape of stoma, number and arrangement of sensory structures on male tail.

Panagrolaimus orthomici Korentchenko, 1992 (Korentchenko, 1992) can be easily separated from BSS8 in having two subdorsal teeth in the stoma (single in BSS8), more anterior position of the excretory pore, different arrangement of sensory structures on male tail.

In conclusion, currently available morphological evidence, along with phylogenetic placement based on UCEs, confirms that the nematode originally identified as *Panagrolaimus* "population II" in (Bostrom, 1988) and subsequently as *Panagrolaimus detritophagus* BSS8 and *Panagrolaimus* BSS8 is in fact *Neocephalobus halophilus* Paetzold, 1958.

4.3 A minimum UCE informative set can accurately classify taxa to genus level

Applying machine learning to UCE presence-absence data provided valuable insights into taxonomic classification within Nematoda. In Rhabditidae, the XGBoost model outperformed all other models (Random Forest, PCA-Logistic Regression, and k-Nearest Neighbors) with an AUC score of 0.9997. Feature selection using the non-zero importance scores derived from XGBoost revealed that 46 UCEs are important for genus-level classification. Also, uce.15118 was the most informative. The optimized XGBoost model was applied to the Panagrolaimidae dataset showing its versatility. The model with the complete set of features obtained accuracy of 82.35%. But, a model only with the top 63 important UCEs obtained perfect classification accuracy (100%). This shows that an efficient classifier for taxonomic assignment can be constructed from a smaller set of informative UCEs, where each feature with non-zero importance was significantly informative.

Through comparative analyses, it was observed that certain UCEs are shared across genera. The most frequently observed UCEs were also important in the XGBoost model. In Panagrolaimidae, the uce.8786, uce.8373, and uce.5498 UCEs appeared in all three categories, indicating their potential role in genus differentiation. Although XGBoost can predict well, it is essential to recognize its limitations. The performance of the model people create will depend upon the quality and completeness of the input data, and any species sampling biases or gaps in the UCE dataset can affect results.

4.4 UCE-Based Nematode Classification potential for biodiversity assessments and agricultural monitoring

Nematodes serve as essential bioindicators of soil quality and ecosystem function, making their rapid and accurate classification essential for environmental monitoring and agricultural research. Their classification

in colonizer-persister categories (cp) as well as the characterization of feeding types based on taxonomic classification (family and genera respectively) have been used for soil health characterization (Du Preez et al., 2022; Ferris, Bongers, & de Goede, 2001; T. Martin, Wade, Singh, & Sprunger, 2022). For instance, opportunistic nematodes (cp-1 and cp-2) reproduce fast and indicate disturbed, nutrient-enriched, or polluted soils, whereas persisters (cp-4 and cp-5) reproduce slowly and indicate stable ecosystems with good soil health. Additionally, the types of feeding and the proportion of different feeding types in soils can be indicative of high organic input, organic matter decomposition, or potential nutrient imbalances. Their rapid identification can enable accurate, near-real-time biodiversity assessments, for instance, through the use of portable sequencing technologies (e.g., Nanopore) and the harvesting of UCEs, as well as model classification.

Such methods can be extended beyond free-living nematodes and applied to the analysis of parasitic nematodes, which are currently a critical target for molecular diagnostics due to their significant impact on global agricultural systems and animal health (Abad et al., 2008; Gang & Hallem, 2016; Strydom, Lavan, Torres, & Heaney, 2023). Particularly focusing on agricultural systems, genera as *Meloidogyne*, *Heterodera*, and *Pratylenchus* are known for causing significant crop losses globally (Khan, 2023). Their early detection is vital for timely management strategies (J. T. Jones et al., 2013; Žibrat et al., 2023). Integrating molecular tools such as ultra-conserved elements (UCEs) with portable sequencing technologies (e.g., Nanopore) has the potential to substantially accelerate the detection of plant-parasitic nematodes (PPNs). Notably, the aforementioned PPN genera are closely related to each other (Thapa, Gates, Reuter-Carlson, Androwski, & Schroeder, 2019), and belong to Clade IV of the Nematoda, alongside the families Panagrolaimidae and Cephalobidae, which were analyzed in this study (M. Blaxter, 2011). This phylogenetic proximity suggests that the bait set developed here could also be effective for targeting UCEs from PPN taxa. Consequently, this approach could serve as a valuable tool to advance both ecological research and applied agricultural diagnostics.

Future advancements in automation could further enhance the speed and scalability of taxonomic classification pipelines, making them even more suitable for large-scale ecological and environmental monitoring. Currently, our approach focuses on taxa within Rhabditidae and Panagrolaimidae; however, we anticipate that extending this method to other ecologically relevant nematode taxa could provide valuable insights, particularly in soil health assessments.

Our study highlights the effectiveness of the newly designed bait sets for multiple loci (UCEs) and target enrichment as a valuable genomic resource for investigating evolutionary relationships among nematodes, specifically within the Rhabditidae and Panagrolaimidae families. This approach enables the retrieval of thousands of genetic markers with known homology, producing results comparable to those obtained using orthologous genes, while allowing for the study of a broader range of taxa at lower costs. While not explored here, the high variability of the flanking region in the targeted conserved loci can allow for fine-scale analysis at the population level, enabling comparisons between the same species occurring at different geographical locations or under different environmental stressors.

5 Author contributions

The study was designed by L.V., L.J. and P.H.S.. J.v.S provided support with bioinformatic pipelines. Laboratory work was conducted by L.V Bioinformatic and statistical work was conducted by L.V. and L.J. P.H.S., A-M.W and O.H. provided supervision during the entire project. The manuscript draft was written by L.V. and L.J., all authors further edited the manuscript..

6 Acknowledgments

This research was conducted under the DFG funded project B08 in the CRC1211 [grant number 268236062] awarded to P. H. Schiffer and A-M Waldvogel, the ENP grant awarded to P. H. Schiffer [grant number: 434028868] and under the Biodiversity Genomics Center Cologne (BioC2) project funded by the UoC forum under the Excellent Research Support Program of the University of Cologne.

7 Conflict of interest statement

The authors declare that they have no conflict of interest.

8 Data availability statement

The sequencing data will be made available through the European Nucleotide Archive (ENA). The bioinformatic pipeline developed for machine learning genera classification is available through CodeOcean (capsule number to be defined).

References

Abad, P., Gouzy, J., Aury, J.-M., Castagnone-Sereno, P., Danchin, E. G. J., Deleury, E., ... Wincker, P. (2008, July). Genome sequence of the metazoan plant-parasitic nematode meloidogyne incognita. *Nature Biotechnology*, *26*(8), 909–915. Retrieved from http://dx.doi.org/10.1038/nbt.1482 doi: 10.1038/nbt.1482

Abolafia, J., & Vecchi, M. (2021, January). Redescription and phylogenetic analysis of the type species of the genus panagrellus thorne, 1938 (rhabditida, panagrolaimidae), p. pycnus thorne, 1938, including the first sem study. *Journal of Nematology*, *53*(1), 1–20. Retrieved from http://dx.doi.org/10.21307/jofnem-2021-080 doi: 10.21307/jofnem-2021-080

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016, January). Harnessing the power of radseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81–92. Retrieved from http://dx.doi.org/10.1038/nrg.2015.28 doi: 10.1038/nrg.2015.28

Andrews, S. (2010). *Fastqc: A quality control tool for high throughput sequence data.* https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed: 2024-04-02)

Andrássy, I. (1984). *Klasse nematoda: (ordnungen monhysterida, desmoscolecida, araeolaimida, chromadorida, rhabditida).* Gustav Fischer Verlag.

Bianchini, G., & Sánchez-Baracaldo, P. (2024, February). ¡scp¿treeviewer¡/scp¿: Flexible, modular software to visualise and manipulate phylogenetic trees. *Ecology and Evolution*, *14*(2). Retrieved from http://dx.doi.org/10.1002/ece3.10873 doi: 10.1002/ece3.10873

Blaimer, B. B., Brady, S. G., Schultz, T. R., Lloyd, M. W., Fisher, B. L., & Ward, P. S. (2015, December). Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. *BMC Evolutionary Biology*, *15*(1). Retrieved from http://dx.doi.org/10.1186/s12862-015-0552-5 doi: 10.1186/s12862-015-0552-5

Blaxter, M. (2011, April). Nematodes: The worm and its relatives. *PLoS Biology*, *9*(4), e1001050. Retrieved from http://dx.doi.org/10.1371/journal.pbio.1001050 doi: 10.1371/journal.pbio.1001050

Blaxter, M. L., De Ley, P., Garey, J. R., Liu, L. X., Scheldeman, P., Vierstraete, A., ... Thomas, W. K. (1998). A molecular evolutionary framework for the phylum nematoda. *Nature*, *392*(6671), 71–75. doi: 10.1038/32160

Bostrom, S. (1988). Descriptions and morphological variability of three populations of panagrolaimus fuchs, 1930 (nematoda: Panagrolaimidae). *Nematologica*, *34*(2), 144–155. Retrieved from http://dx.doi.org/10.1163/002825988X00233 doi: 10.1163/002825988x00233

Bushnell, B. (2014). *Bbmap: A fast, accurate, splice-aware aligner.* https://sourceforge.net/projects/bbmap/. (Accessed: 2024-04-02)

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, *abs/1603.02754*. Retrieved from http://arxiv.org/abs/1603.02754

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., ... Li, H. (2021, January). Twelve years of samtools and bcftools. *GigaScience*, *10*(2). Retrieved from http://dx.doi.org/10.1093/gigascience/giab008 doi: 10.1093/gigascience/giab008

Duckett, D. J., Calder, K., Sullivan, J., Tank, D. C., & Carstens, B. C. (2023, November). Reduced representation approaches produce similar results to whole genome sequencing for some common phylogeographic analyses. *PLOS ONE*, *18*(11), e0291941. Retrieved from http://dx.doi.org/10.1371/journal.pone.0291941 doi: 10.1371/journal.pone.0291941

Du Preez, G., Daneel, M., De Goede, R., Du Toit, M. J., Ferris, H., Fourie, H., ... Schmidt, J. H. (2022, June). Nematode-based indices in soil ecology: Application, utility, and future directions. *Soil Biology and Biochemistry*, *169*, 108640. Retrieved from http://dx.doi.org/10.1016/j.soilbio.2022.108640 doi: 10.1016/j.soilbio.2022.108640

Erickson, K. L., Pentico, A., Quattrini, A. M., & McFadden, C. S. (2020, September). New approaches to species delimitation and population structure of anthozoans: Two case studies of octocorals using ultraconserved elements and exons. *Molecular Ecology Resources*, *21*(1), 78–92. Retrieved from http://dx.doi.org/10.1111/1755-0998.13241 doi: 10.1111/1755-0998.13241

Faircloth, B. C. (2015, November). Phyluce is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*(5), 786–788. Retrieved from http://dx.doi.org/10.1093/bioinformatics/btv646 doi: 10.1093/bioinformatics/btv646

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012, January). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, *61*(5), 717–726. Retrieved from http://dx.doi.org/10.1093/sysbio/SYS004 doi: 10.1093/sysbio/sys004

Felix, M.-A., De Ley, P., Sommer, R. J., Frisse, L., Nadler, S. A., Thomas, W. K., ... Sternberg, P. W. (2000). Evolution of vulva development in the cephalobina (nematoda). *Developmental Biology*, *221*, 68–86.

Ferris, H., Bongers, T., & de Goede, R. (2001, September). A framework for soil food web diagnostics: extension of the nematode faunal analysis concept. *Applied Soil Ecology*, *18*(1), 13–29. Retrieved from http://dx.doi.org/10.1016/S0929-1393(01)00152-4 doi: 10.1016/s0929-1393(01)00152-4

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. Retrieved from https://www.jstatsoft.org/index.php/jss/article/view/v033i01 doi: 10.18637/jss.v033.i01

Gang, S. S., & Hallem, E. A. (2016, July). Mechanisms of host seeking by parasitic nematodes. *Molecular and Biochemical Parasitology*, *208*(1), 23–32. Retrieved from http://dx.doi.org/10.1016/j.molbiopara.2016.05.007 doi: 10.1016/j.molbiopara.2016.05.007

Gilbert, P. S., Chang, J., Pan, C., Sobel, E. M., Sinsheimer, J. S., Faircloth, B. C., & Alfaro, M. E. (2015, November). Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. *Molecular Phylogenetics and Evolution*, *92*, 140–146. Retrieved from http://dx.doi.org/10.1016/j.ympev.2015.05.027 doi: 10.1016/j.ympev.2015.05.027

Guyon, I., & Elisseeff, A. (2003, March). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, *3*(null), 1157–1182.

Hodda, M. (2022, March). Phylum nematoda: a classification, catalogue and index of valid genera, with a census of valid species. *Zootaxa*, *5114*(1), 1–289.

Jones, J. T., Haegeman, A., Danchin, E. G. J., Gaur, H. S., Helder, J., Jones, M. G. K., ... Perry, R. N. (2013, July). Top 10 plant-parasitic nematodes in molecular plant pathology. *Molecular Plant Pathology*, *14*(9), 946–961. Retrieved from http://dx.doi.org/10.1111/mpp.12057 doi: 10.1111/mpp.12057

Jones, M. R., & Good, J. M. (2015, July). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*(1), 185–202. Retrieved from http://dx.doi.org/10.1111/mec.13304 doi: 10.1111/mec.13304

Khan, M. R. (2023). Nematode pests of agricultural crops, a global overview. In *Novel biological and biotechnological applications in plant nematode management* (p. 3–45). Springer Nature Singapore. Retrieved from http://dx.doi.org/10.1007/978-981-99-2893-4_1 doi: 10.1007/978-981-99-2893-4_1

Korentchenko, E. A. (1992). Panagrolaimus orthomici sp.n. (cephalobina), a nematode of bark beetles of the genus panagrolaimus in north-eastern asia. *Parazitologia*, *6*, 530–534.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. Retrieved from https://www.jstatsoft.org/index.php/jss/article/view/v028i05 doi: 10.18637/jss.v028.i05

Kuhn, R. M., Haussler, D., & Kent, W. J. (2012, August). The ucsc genome browser and associated tools. *Briefings in Bioinformatics*, *14*(2), 144–161. Retrieved from http://dx.doi.org/10.1093/bib/bbs038 doi: 10.1093/bib/bbs038

Lewis, S. C., Dyal, L. A., Hilburn, C. F., Weitz, S., Liau, W.-S., LaMunyon, C. W., & Denver, D. R. (2009, January). Molecular evolution in panagrolaimus nematodes: origins of parthenogenesis, hermaphroditism and the antarctic species p. davidi. *BMC Evolutionary Biology*, *9*(1). Retrieved from http://dx.doi.org/10.1186/1471-2148-9-15 doi: 10.1186/1471-2148-9-15

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from http://CRAN.R-project.org/doc/Rnews/

Libbrecht, M. W., & Noble, W. S. (2015, June). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.*, *16*(6), 321–332.

Lu, Y., Li, M., Gao, Z., Ma, H., Chong, Y., Hong, J., ... Deng, W. (2025, January). Advances in whole genome sequencing: Methods, tools, and applications in population genomics. *International Journal of Molecular Sciences*, *26*(1), 372. Retrieved from http://dx.doi.org/10.3390/ijms26010372 doi:

10.3390/ijms26010372

Lunter, G., & Goodson, M. (2010, October). Stampy: A statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome Research*, *21*(6), 936–939. Retrieved from http://dx.doi.org/10.1101/gr.111120.110 doi: 10.1101/gr.111120.110

Lv, Z., Li, M., Wang, Y., & Zou, Q. (2023, February). Editorial: Machine learning for biological sequence analysis. *Front. Genet.*, *14*, 1150688.

Martin, M. (2011, May). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10. Retrieved from http://dx.doi.org/10.14806/ej.17.1.200 doi: 10.14806/ej.17.1.200

Martin, T., Wade, J., Singh, P., & Sprunger, C. D. (2022, March). The integration of nematode communities into the soil biological health framework by factor analysis. *Ecological Indicators*, *136*, 108676. Retrieved from http://dx.doi.org/10.1016/j.ecolind.2022.108676 doi: 10.1016/j.ecolind.2022.108676

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020, February). Iq-tree 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. Retrieved from http://dx.doi.org/10.1093/molbev/msaa015 doi: 10.1093/molbev/msaa015

Natsidis, P., Kapli, P., Schiffer, P. H., & Telford, M. J. (2021, February). Systematic errors in orthology inference and their effects on evolutionary analyses. *iScience*, *24*(2), 102110.

Paetzold, R. (1958). Neocephalobus halophilus n. sp. (nematoda: Cephalobidae) aus salzböden der umgebung von halle (saale). *Archiv für Naturgeschichte*, *22*(2–3), 157–167.

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*, 526-528. doi: 10.1093/bioinformatics/bty633

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020, June). Using spades de novo assembler. *Current Protocols in Bioinformatics*, *70*(1). Retrieved from http://dx.doi.org/10.1002/cpbi.102 doi: 10.1002/cpbi.102

Qing, X., Zhang, Y. M., Sun, S., Ahmed, M., Lo, W.-S., Bert, W., ... Li, H. (2024, December). Phylogenomic insights into the evolution and origin of nematoda. *Systematic Biology*. Retrieved from http://dx.doi.org/10.1093/sysbio/syae073 doi: 10.1093/sysbio/syae073

Quattrini, A. M., Faircloth, B. C., Dueñas, L. F., Bridge, T. C. L., Brugler, M. R., Calixto-Botía, I. F., ... McFadden, C. S. (2017, December). Universal target-enrichment baits for anthozoan (cnidaria) phylogenomics: New approaches to long-standing problems. *Molecular Ecology Resources*, *18*(2), 281–295. Retrieved from http://dx.doi.org/10.1111/1755-0998.12736 doi: 10.1111/1755-0998.12736

Quinlan, A. R., & Hall, I. M. (2010, January). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. Retrieved from http://dx.doi.org/10.1093/bioinformatics/btq033 doi: 10.1093/bioinformatics/btq033

Revell, L. J. (2024, January). phytools 2.0: an updated r ecosystem for phylogenetic comparative methods (and other things). *PeerJ*, *12*, e16505. Retrieved from http://dx.doi.org/10.7717/peerj.16505 doi: 10.7717/peerj.16505

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011, March). pROC: an open-source package for R and s+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77.

Rubin, B. E. R., Ree, R. H., & Moreau, C. S. (2012, April). Inferring phylogenies from rad sequence data. *PLoS ONE*, *7*(4), e33394. Retrieved from http://dx.doi.org/10.1371/journal.pone.0033394 doi: 10.1371/journal.pone.0033394

Saeys, Y., Inza, I., & Larrañaga, P. (2007, 08). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507-2517. Retrieved from https://doi.org/10.1093/bioinformatics/btm344 doi: 10.1093/bioinformatics/btm344

Schiffer, P. H., Nsah, N. A., Grotehusmann, H., Kroiher, M., Loer, C., & Schierenberg, E. (2014, May). Developmental variations among panagrolaimid nematodes indicate developmental system drift within a small taxonomic unit. *Development Genes and Evolution*, *224*(3), 183–188. Retrieved from http://dx.doi.org/10.1007/s00427-014-0471-2 doi: 10.1007/s00427-014-0471-2

Shannon, A. J., Browne, J. A., Boyd, J., Fitzpatrick, D. A., & Burnell, A. M. (2005, June). The an-

hydrobiotic potential and molecular phylogenetics of species and strains ofpanagrolaimus(nematoda, panagrolaimidae). *Journal of Experimental Biology*, *208*(12), 2433–2445. Retrieved from http://dx.doi.org/10.1242/jeb.01629 doi: 10.1242/jeb.01629

Shatilovich, A., Gade, V. R., Pippel, M., Hoffmeyer, T. T., Tchesunov, A. V., Stevens, L., ... Kurzchalia, T. V. (2023, July). A novel nematode species from the siberian permafrost shares adaptive mechanisms for cryptobiotic survival with c. elegans dauer larva. *PLOS Genetics*, *19*(7), e1010798. Retrieved from http://dx.doi.org/10.1371/journal.pgen.1010798 doi: 10.1371/journal.pgen.1010798

Shokoohi, E., & Masoko, P. (2024, November). Morphological and molecular analysis of propanagrolaimus siweyae sp. n. (nematoda: Panagrolaimidae) from molepo dam, limpopo province, south africa, and its relationship with water parameters. *Biologia*, *79*(12), 3573–3587. Retrieved from http://dx.doi.org/10.1007/s11756-024-01806-2 doi: 10.1007/s11756-024-01806-2

Steiner, G. (1929a). Cephalobus (neocephalobus) aberrans n. sg. n. sp., (rhabditidae, nematodes) from the feces of a guinea-pig. *Journal of Parasitology*, *16*(2), 88–90. Retrieved from https://www.jstor.org/stable/3271914 doi: 10.2307/3271914

Steiner, G. (1929b). Cephalobus (neocephalobus) aberrans n.sg., n.sp. (rhabditidae, nematodes) from the feces of a guinea-pig. *Journal of Parasitology*, *26*, 88–90.

Stevens, L., Félix, M.-A., Beltran, T., Braendle, C., Caurcel, C., Fausett, S., ... Blaxter, M. (2019, April). Comparative genomics of 10 newcaenorhabditisspecies. *Evolution Letters*, *3*(2), 217–236. Retrieved from http://dx.doi.org/10.1002/evl3.110 doi: 10.1002/evl3.110

Strydom, T., Lavan, R. P., Torres, S., & Heaney, K. (2023, May). The economic impact of parasitism from nematodes, trematodes and ticks on beef cattle production. *Animals*, *13*(10), 1599. Retrieved from http://dx.doi.org/10.3390/ani13101599 doi: 10.3390/ani13101599

Thapa, S., Gates, M. K., Reuter-Carlson, U., Androwski, R. J., & Schroeder, N. E. (2019, March). Convergent evolution of saccate body shapes in nematodes through distinct developmental mechanisms. *EvoDevo*, *10*(1). Retrieved from http://dx.doi.org/10.1186/s13227-019-0118-5 doi: 10.1186/s13227-019-0118-5

van den Hoogen, J., Geisen, S., Wall, D. H., Wardle, D. A., Traunspurger, W., de Goede, R. G. M., ... Crowther, T. W. (2020, March). A global database of soil nematode abundance and functional group composition. *Scientific Data*, *7*(1). Retrieved from http://dx.doi.org/10.1038/s41597-020-0437-3 doi: 10.1038/s41597-020-0437-3

van der Sprong, J., de Voogd, N. J., McCormack, G. P., Sandoval, K., Schätzle, S., Voigt, O., ... Vargas, S. (2023, November). A novel target-enriched multilocus assay for sponges (porifera): Red sea haplosclerida (demospongiae) as a test case. *Molecular Ecology Resources*, *24*(2). Retrieved from http://dx.doi.org/10.1111/1755-0998.13891 doi: 10.1111/1755-0998.13891

Van Megen, H., van den Elsen, S., Holterman, M., Karssen, G., Mooyman, P., Bongers, T., ... Helder, J. (2009). A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal dna sequences. *Nematology*, *11*(6), 927–950. doi: 10.1163/156854109X456862

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer. Retrieved from https://www.stats.ox.ac.uk/pub/MASS4/ (ISBN 0-387-95457-0)

Villegas, L., Pettrich, L., Acevedo-Trejos, E., Suwanngam, A., Wassey, N., Allende, M. L., ... Schiffer, P. H. (2024, October). Hierarchical patterns of soil biodiversity in the atacama desert: Insights across biological scales. Retrieved from http://dx.doi.org/10.1101/2024.09.30.615889 doi: 10.1101/2024.09.30.615889

Williams, M. (1986). The use of scanning electron microscopy in the taxonomy of panagrolaimus (nematoda: Panagrolaimidae). *Nematologica*, *32*(1), 89–97. Retrieved from http://dx.doi.org/10.1163/187529286X00057 doi: 10.1163/187529286x00057

Winker, K., Glenn, T. C., & Faircloth, B. C. (2018, October). Ultraconserved elements (uces) illuminate the population genomics of a recent, high-latitude avian speciation event. *PeerJ*, *6*, e5735. Retrieved from http://dx.doi.org/10.7717/peerj.5735 doi: 10.7717/peerj.5735

Žibrat, U., Viaene, N., Širca, S., van Beek, J., Susič, N., & Stare, B. G. (2023, December). Nemdetect: Early detection of quarantine nematodes in potatoes using remote sensing. *EFSA Supporting Publications*, *20*(12). Retrieved from https://data.europa.eu/doi/10.2903/sp.efsa.2023.EN-8143 doi: 10.2903/sp.efsa.2023.EN-8143