Identification, evolutionary history and characteristics of orphan genes in root-knot nematodes

Ercan Seçkin[1,2] , Dominique Colinet[1] , Marc Bailly-Bechet[1], Aurélie Seassau[1], Silvia Bottini[1], Edoardo Sarti[2], Etienne GJ Danchin[1]

1. Institut Sophia Agrobiotech, INRAE, Université Côte d'Azur, CNRS, Sophia-Antipolis 06903, France
2. Centre Inria at Université Côte d'Azur, Sophia-Antipolis 06902, France

Abstract

Orphan genes, lacking homologs in other species, are systematically found across genomes. Their presence may result from extensive divergence from pre-existing genes or from *de novo* gene birth, which occurs when a gene emerges from a previously non-genic region. In this study, we identified orphan genes in the genomes of globally distributed plant-parasitic nematodes of the genus *Meloidogyne* and investigated their origins, evolution, and characteristics. Using a comparative genomics framework across 85 nematode species, we found that 18% of *Meloidogyne* genes are genus-specific, transcriptionally supported orphans. By combining ancestral sequence reconstruction and synteny-based approaches, we inferred that 20% of these orphan genes originated through high divergence, while 18% likely emerged *de novo*. Proteomic and translatomic evidence confirmed the translation of a subset of these genes, and feature analyses revealed distinctive molecular signatures, including shorter length, signal peptide enrichment, and a tendency for extracellular localization. These findings highlight orphan genes as a substantial and previously underexplored component of the *Meloidogyne* genome, with potential roles in their worldwide parasitism.

Introduction

Orphan genes and *de novo* gene birth

In eukaryotes, genomes systematically contain a number of genes (up to 30%) that have no identifiable homolog in other species, known as orphan genes (Tautz and Domazet-Lošo 2011; Seçkin, Colinet, Sarti, and E.G.J. Danchin 2025). The emergence of orphan genes is an important evolutionary mechanism that can be associated with the acquisition of functional novelty. Orphan genes may originate from a pre-existing gene that has accumulated high divergence reaching the point of no recognizable homology. This process can be facilitated by gene duplication or horizontal gene transfer events, followed by rapid evolution. Studies in several species suggest, however, that this explanation applies only to a subset of existing orphan genes (Vakirlis et al. 2020). Another non-mutually exclusive hypothesis is that orphan

genes may emerge from non-genic regions. This phenomenon, known as *de novo* gene birth, occurs when previously non-coding and/or non-transcribed DNA sequences acquire the capacity to be coding/ transcribed then translated to a functional protein (Schmitz and Bornberg-Bauer 2017; Weisman 2022). For a long time, *de novo* emergence was considered highly unlikely as (i) it was believed that a new gene can only emerge from an existing gene and (ii) the probability of a newly emerged gene coding for a functional protein to be maintained in populations by selection is considered extremely low (Jacob 1977). However, with the explosion of genomic sequencing projects and the resulting increase in available genome data from a higher diversity of species, it was realized that *de novo* gene emergence is not as rare as initially thought, and that many species- or lineage-specific sequences lack recognizable homologs (Khalturin et al. 2009). Several studies have taken advantage of this richer set of genome data to confirm the existence of genes that likely emerged *de novo* (Tautz and Domazet-Lošo 2011; McLysaght and Hurst 2016; Van Oss and Carvunis 2019).

In the case of a protein-coding gene, *de novo* emergence involves two main distinct processes: (i) transcription of initially non-coding DNA, and (ii) acquisition of an open reading frame (ORF). The order of these events allows two mechanisms to be distinguished (Van Oss and Carvunis 2019): "transcription first" and "ORF first". As their names suggest, these mechanisms differ in the order between transcription and ORF acquisition. However, it is important to note that the distinction between the "transcription first" and "ORF first" mechanisms is not always straightforward. Just as it can be difficult to definitively classify an orphan gene as *de novo* or highly diverged, the temporal sequence of transcription and ORF acquisition may not be neatly separated. For instance, an ORF formed in a region of low transcription can gradually acquire regulatory features. Alternatively a *de novo* gene can later undergo rapid divergence, which can obscure its origin.

Methods to identify orphan genes and *de novo* gene birth

The most common approach to identifying orphan genes is to start from a focal branch in the tree of life and search for homologs in other species using comparative genomics. One widely used method is phylostratigraphy, which aims at identifying homologs for each gene from a species or clade of interest using BLAST (Camacho et al. 2009) or similar tools against other species. It should be noted that most of these tools use protein sequences as a proxy for protein-coding genes. Then, based on these searches, groups or clusters of homologous genes are built using state-of-the-art software such as OrthoFinder (Emms and Kelly 2019), ORFan-Finder (Ekstrom and Yin 2016) or SonicParanoid (Cosentino and Iwasaki 2023). Exclusive identification of a gene within one or a few closely related species enables determination of probable relative timing of gene emergence, as well as classification of the gene as orphan.

From an initial dataset of orphan genes, *de novo* genes can be identified by aligning the corresponding proteins to the genome of the most closely related species translated into 6 frames and looking for similarities in the corresponding non-coding regions. If the

corresponding region in the related species is non-coding and mutations can be identified at specific positions that have led to the acquisition of an ORF, then a *de novo* emergence event can be assumed. However, establishing reliable correspondences between genomes can be difficult in case of high divergence. Translocations, structural changes, and incomplete assemblies can also obscure the ancestral origin of a gene. Distinguishing between *de novo* gene birth and high divergence from homologs is particularly challenging because highly diverged homologs no longer have detectable sequence similarity, making them appear to have arisen *de novo*. Conversely, *de novo* genes arise from non-coding sequences that may, through convergent evolution, superficially resemble highly divergent homologs, further complicating their identification.

Nevertheless, incorporating the broader genomic context via conserved synteny analysis can help distinguish between these two possibilities. This involves determining if the genes surrounding the candidate orphan gene in the focal species are colinearly conserved in closely related target species. If the surrounding genes are conserved, the next step is to examine the homologous target locus corresponding to the candidate orphan gene. If another gene is present at this target locus but lacks homology to the orphan gene, we can hypothesize that the orphan gene has highly diverged from the ancestral gene, although translocation of a gene at this locus must also be considered. Conversely, if there is no predicted gene at this locus but there is partial alignment of the orphan gene with frameshifts and/or invalid splice sites, then the *de novo* gene birth hypothesis is more likely.

For more details, a recent paper reviews the challenges and limitations of methods to detect *de novo* gene emergence (Grandchamp et al. 2025).

Studies of orphan and *de novo* genes in living

Orphan genes have been identified in many species, and their proportion in the genome varies according to the species (Seçkin, Colinet, Sarti, and E. Danchin 2025). For instance, orphan genes represent 1.3% of the human genes (Toll-Riera et al. 2008), 2.3% of the *Drosophila melanogaster* genes (Wissler et al. 2013), 4.4% of the *Caenorhabditis elegans* genes (Zhang et al. 2019), 25% of the *Saccharomyces cerevisiae* genes (Dujon 1996), and 33% of the *Pristionchus pacificus* genes, which is a nematode closely related to *C. elegans* (Prabh and Rödelsperger 2019).

The *de novo* emergence of orphan genes has been described in several species (Van Oss and Carvunis 2019). This phenomenon was first described for *D. melanogaster* in 2006 (Levine et al. 2006) and for *S. cerevisiae* in 2008 (Cai et al. 2008). Subsequently, it was shown that 12% of orphan genes in *D. melanogaster* (Zhou et al. 2008) and 5.5% of orphan genes in humans (Toll-Riera et al. 2008) have probably emerged *de novo*. However, these figures seem low in light of a more recent study that estimates most orphan genes emerged *de novo* and suggests that there are *de novo* genes yet to be identified in these species (Vakirlis et al. 2020). Little is known about the function of *de novo* genes. In yeast, a study showed that a *de novo* gene encodes a protein involved in the DNA repair pathway during the stationary phase of *S.*

cerevisiae. This protein contributes to the yeast's robustness when placed in a nutrient-poor environment (Cai et al. 2008). In plants, the first *de novo* gene whose function has been characterized is QQS, an *Arabidopsis thaliana* gene identified in 2009 that is involved in carbon and nitrogen metabolism (Silveira et al. 2013). In codfish, a *de novo* gene encoding an antifreeze glycoprotein has been identified (Zhuang et al. 2019). Finally, a study has identified a human-specific de novo gene encoding a protein found in the brain that appears to be associated with Alzheimer's disease (Li et al. 2010). In a recent review, it has been observed that many orphan genes in animals, and more specifically *de novo* ones, are involved in reproductive functions or are specifically expressed in gonads (Seçkin, Colinet, Sarti, and E. Danchin 2025).

Nematodes of the genus *Meloidogyne*

Nematodes are arguably the most numerous animals on Earth, only paralleled by arthropods and found in every biotope (Eisenback and Triantaphyllou 2020). In addition to their prevalence, they comprise free-living species as well as animal and plant parasites which cause substantial health problems and have important economic impact. So far, studies of orphan genes in nematodes have been restricted to the model species *C. elegans* and *P. pacificus*, which are both free-living species and relatively closely related. Despite their importance, orphan genes have never been investigated in parasitic nematodes.

Overall, more than 27,000 nematode species are described, and grouped into 12 main clades (Van Megen et al. 2009). Of these, four contain plant parasites, (Holterman et al. 2017; Ahmed et al. 2022; Qing et al. 2024), which collectively cause over $157 billion of economic loss every year to the global agricultural production (Lilley et al. 2024). The most problematic of these worms regarding agriculture are the root-knot nematodes of the genus *Meloidogyne*, which belongs to clade 12 (*Tylenchida*) (Ahmed et al. 2022). In addition to their economic importance, these species display peculiar biological features such as diverse reproductive modes and ploidy levels. *Meloidogyne* species have been classified in three main clades (I-III), based on ribosomal DNA markers (De Ley et al. 2002). Most species in clade I are polyploid and reproduce exclusively via parthenogenesis while in the other clades, most species are diploid and able to switch between parthenogenesis and sexual reproduction.

Meloidogyne species hijack plant root development to induce giant cells from which they feed, leaving galls symptoms on plant roots and causing billions of dollars in damage to agricultural production each year. Their life cycle, lasting three to six weeks, is divided into two phases: an exophytic phase outside the plant and an endophytic phase within the plant roots. The exophytic phase begins with the J1 juvenile larval form (in the egg), and is followed by the J2 stage, responsible for penetrating the root and initiating the endophytic phase. This phase includes the J3 and J4 larval stages, which are difficult to distinguish, followed by the transition of the larva to the adult stage. Adult males exit the root, while females remain sedentary, producing hundreds of eggs that are deposited on the root surface.

The study of *Meloidogyne* genomes began in 2008 with the sequencing of *Meloidogyne incognita* genome (Abad et al. 2008). This was the first genome of a plant-parasitic animal to be sequenced. At that time, only two other annotated nematode genome assemblies were available (*C. elegans* and *Brugia malayi*). Comparisons with those nematode genomes, as well as with the *Drosophila* genome, revealed that approximately half of the protein-coding genes in *M. incognita* had no identifiable homologs in the other species. These genes were considered specific to *M. incognita*. However, the limited number of available nematode genomes at that time likely overestimated this proportion, and no hypothesis was given concerning their origin. More complete and better-annotated versions of the *M. incognita* genome were subsequently obtained, as well as genomes of other nematodes. In 2013, a comparison of predicted genes in *M. incognita* and *M. hapla* with a comprehensive dataset of other species' genomes revealed approximately 10,000 *Meloidogyne* genes that lacked homology in the other species (Danchin et al. 2013a). Once again, the origin of these genes, considered as orphans, was not investigated.
More recently, a comparison of the gene sets of seven *Meloidogyne* species against those of 56 other nematodes, followed by a homology search against the NCBI's non-redundant (nr) library, revealed tens of thousands of proteins specific to *Meloidogyne* species in *M. incognita* (Grynberg et al. 2020a). Most of these genes were supported by transcriptomic data. While the origin of these numerous orphan genes was questioned, it was not investigated whether they highly diverged from ancestral sequences or emerged *de novo*.

Overall, these preliminary studies hypothesized that these numerous lineage-specific genes, supported by transcriptomic data, might play important roles, including for plant -parasitism. Consistent with this view, these nematodes secrete proteins into host plants that act as effectors of parasitism, most of which are encoded by genes lacking identifiable homologs, i.e, orphan genes (Mitchum et al. 2013a). Orphan effector genes involved in parasitic success have attracted a lot of attention as good targets for more specific nematode control. However, despite their importance, the origin, evolutionary history, and prevalence of orphan genes in *Meloidogyne* genomes remain unstudied.

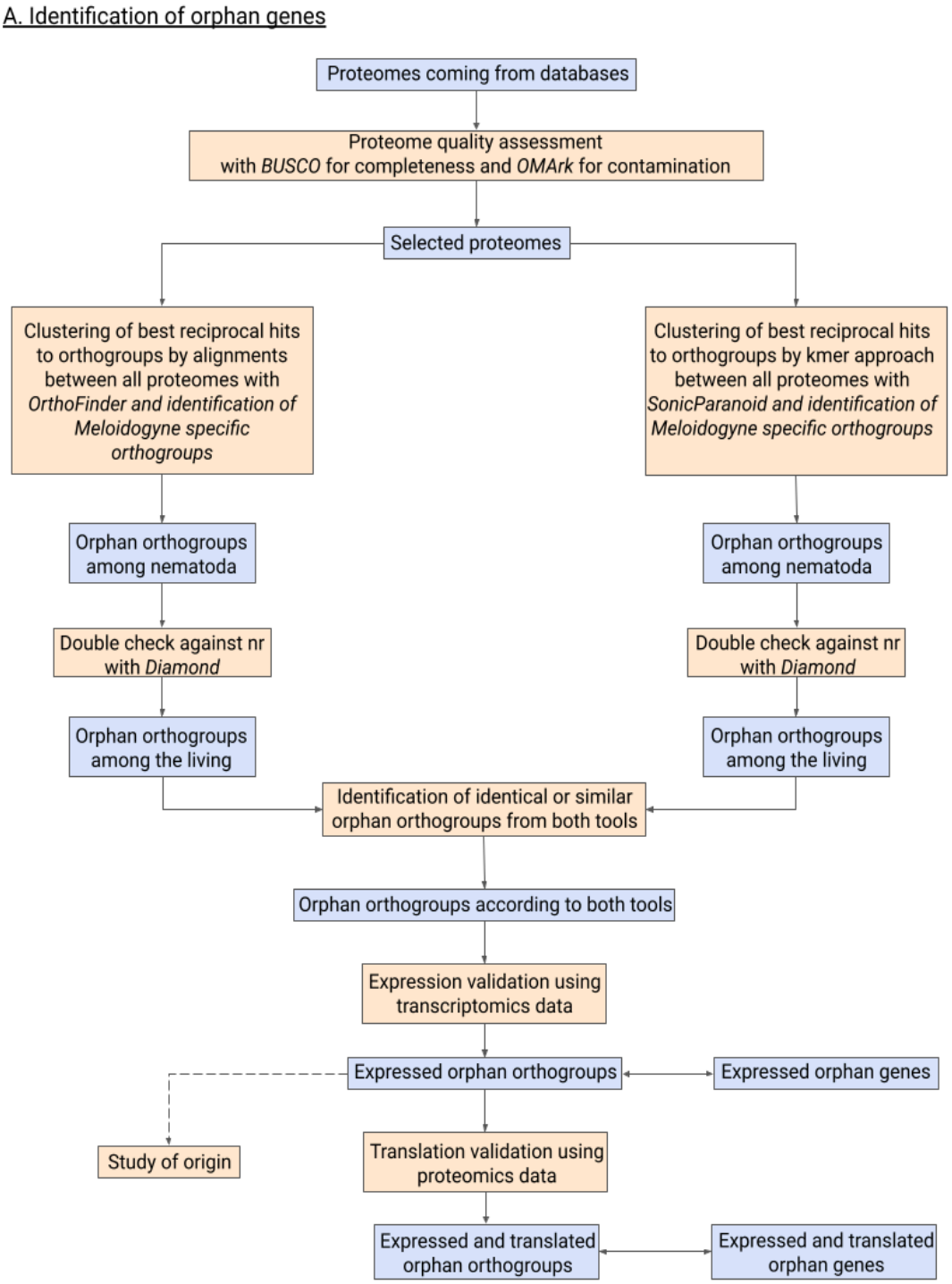
Considering their important economic impact, their peculiar biology and the previous report of numerous candidate orphan genes, it is timely to study the evolutionary origin, fate, characteristics of orphan genes in these species as well as their impact on their genomes and biology.

In this study, we identified orphan genes in eight *Meloidogyne* species, including the most economically important ones (*M. arenaria, M. chitwoodi, M. enterolobii, M. graminicola, M. hapla, M. incognita, M. javanica,* and *M. luci)*. This dataset covers the three main *Meloidogyne* clades and spans an estimated ~65 My of evolution since their last common ancestor (Qing et al. 2025). By comparing their sets of predicted proteins with those of other nematodes and proteins in the NCBI's nr database, we identified proteins lacking homology in other species to date. We then only retained proteins whose corresponding genes were supported by transcriptomic data. Using conserved synteny analysis with the closest available

outgroup species, we further explored the evolutionary history of these orphan genes. Finally, we determined whether orphan genes and proteins differed in some characteristics compared to the rest of the genes in the genomes.

Materials & Methods

A diagram summarizing the entire workflow is shown in Figure 1.



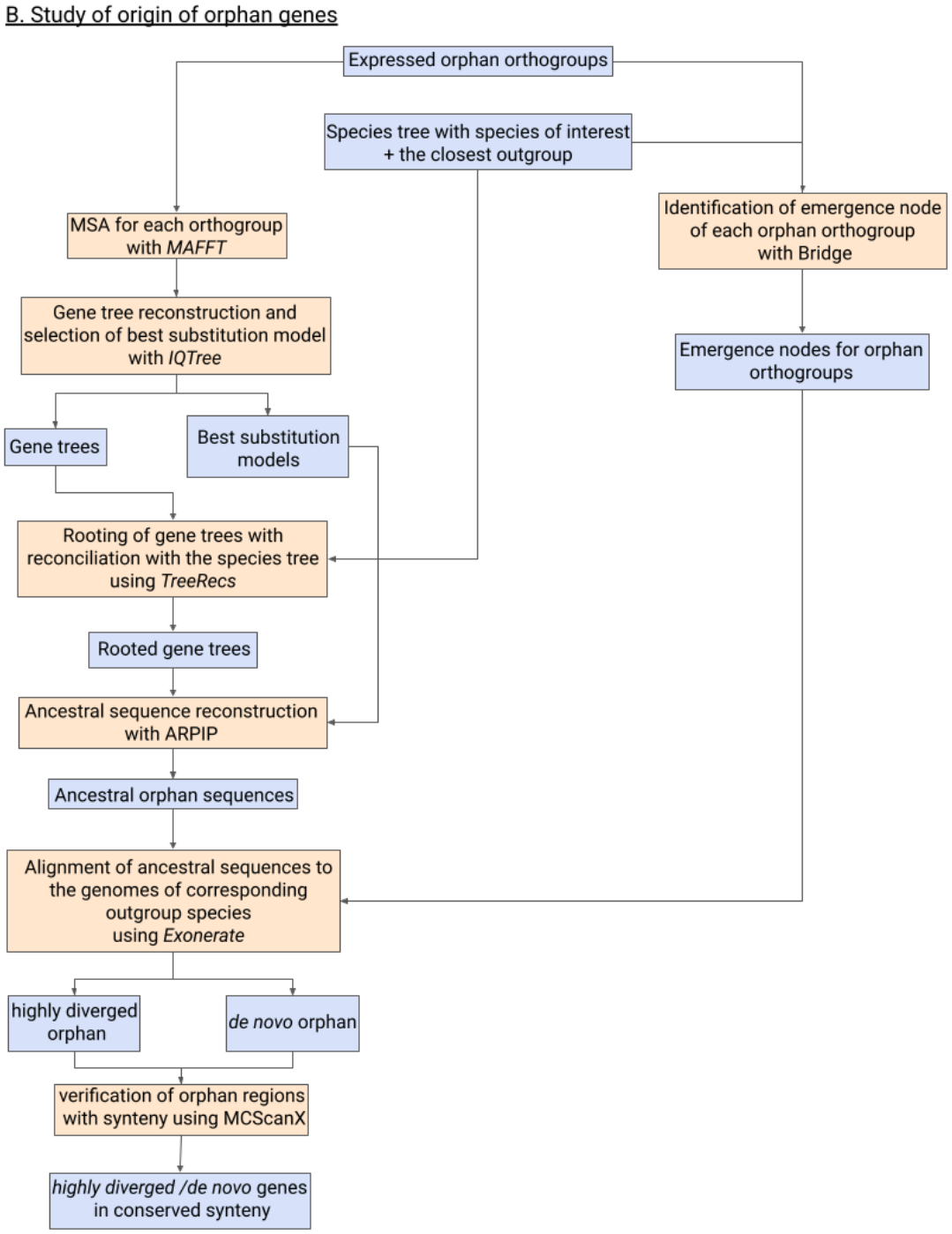


Figure 1: Bioinformatics pipeline to identify orphan genes and to study their evolution in *Meloidogyne*. Representation of different steps of the built pipeline is shown in order. Blue cases correspond to input or outputs and orange cases correspond to different steps or analysis. A. First part of the built pipeline which is used to identify orphan genes in *Meloidogyne*. B. Second part of the pipeline to study the origin of identified orphan genes and to classify orphan genes as highly diverged orphans or de novo emerged ones.

1. Selection of nematode predicted proteomes

A total of 206 nematode predicted proteomes are available in the WormBase Parasite database, including different *Meloidogyne* species (Howe et al. 2017). We made a selection from all proteomes based on two criteria: (i) quality of annotation and (ii) diversity of species represented. To measure the quality of the proteomes, completeness was assessed using BUSCO (2) (v 5.4.4, with metazoa_odb10 as dataset). This tool calculates the percentage of 'universally' conserved genes in a branch of the phylogenetic tree (here *metazoa*) within each proteome and yields a BUSCO score corresponding to the percentage of conserved genes actually found in the proteome. Absence of contamination in the selected proteomes was then assessed with OMArk (Nevers et al. 2022) that calculates possible contaminations on the proteome by flagging a protein as contaminant if a protein has high similarity to sequences from non-target taxa. We defined a minimal BUSCO completeness threshold of 62% and a maximum OMArk contamination of 5%. Concerning the criteria of diversity of species represented, all species of the genus *Meloidogyne*, which is the group of interest, were retained if their proteome satisfied the BUSCO and OMArk criteria for completeness and lack of contamination. For the other genera, a maximum of three species for which the proteome met the above quality criteria were retained for each genus to provide a balanced representation in the dataset of selected proteomes.

2. Comparative analysis of predicted proteomes in Nematoda

Two different tools were used to perform the comparative analysis of the 85 nematode proteomes that were selected in the previous step. One of these tools is OrthoFinder (Emms and Kelly 2019) (v 2.5.4) that performs a homology search of all proteomes against all proteomes using Diamond (Buchfink et al. 2015), a tool similar to BLAST (Camacho et al. 2009) but much faster due to its use of a simplified protein alphabet and binary data to align sequences. OrthoFinder was utilized with default parameters, using dendroblast (Kelly and Maini 2013) for gene tree inference and setting an e-value of 0.001. Based on this homology search between all provided protein sequences, OrthoFinder groups the best mutual hits into orthogroups. For each orthogroup, OrthoFinder generates a phylogenetic tree to identify orthologs. The proteome of the tardigrade *Hypsibius exemplaris* was used as a nematode outgroup to help root the OrthoFinder-reconstructed trees. The other tool used was SonicParanoid2 (Cosentino and Iwasaki 2023), which is analogous in its methodology but uses a different clustering method, using graphs to identify homologs rather than MCL clustering and phylogenies, as OrthoFinder does. Also, SonicParanoid uses MMSeqs in sensitive mode to align sequences all against all using k-mers (Steinegger and Söding 2017). These two tools were chosen mainly to have a range of alignment strategies, one being based on BLAST-like heuristic method and the other utilizing k-mers.

OrthoFinder and SonicParanoid both output two main files. The first file is a matrix of protein counts per species for each orthogroup. The second file contains the orthogroup identifiers and the proteins that belong to these groups. Using a python script, orthogroups containing exclusively proteins from *Meloidogyne* species were initially selected for both

tools. To reduce the risk of including proteins that would be false-predictions of annotation software, all orthogroups comprising a single *Meloidogyne* species (or a single predicted protein-coding orphan gene) were isolated from the orphan dataset. Therefore, the final set of orthogroups that were retained contained proteins present in at least two *Meloidogyne* species and lacked homologs in the other species.

After the homology search (see below), the orthogroups of both tools were compared. We identified identical, similar, and subset orthogroups using a home-made script that compares the elements of each orthogroup between them. Identical orthogroups contained exactly the same proteins in both tools and were retained as such. Similar orthogroups shared at least 75% of their proteins with another orthogroup. For these, a union merged orthogroup was constructed. Finally, subset orthogroups contained only a portion of the proteins from an orthogroup identified by the other tool. In such cases, the largest orthogroup was retained. All retained proteins were considered encoded by potential *Meloidogyne*-specific orphan genes lacking homology in other nematodes.

3. Homology search among living organisms

A homology search was performed with the above-mentioned *Meloidogyne*-species orphan proteins using Diamond (Buchfink et al. 2015) against the NCBI non-redundant (nr) protein database. The e-value was set to 0.001 and the minimum percentage of identity between two sequences to 25%. The more-sensitive option was used to enhance the sensitivity of the analysis and obtain results that are close to those of BLASTp. The top100 hits were used to infer a taxonomic annotation. Diamond was used in last common ancestor (LCA) mode to assign taxonomic labels to the sequences from the NCBI taxonomic tree. The protein list for each *Meloidogyne* species was obtained, along with the taxon ID of the last common ancestor of their hits for each protein query. A homemade python script was used with the ete3 package (Huerta-Cepas et al. 2016) to convert taxon ID numbers into taxon names. Next, we eliminated orthogroups that contained at least one protein having a significant hit outside of *Meloidogyne*.

4. Transcriptional support for orphan genes

Short-read paired-end RNA-seq (Wang et al. 2009) transcriptomic data were available for *M. incognita* (Blanc-Mathieu et al. 2017) (data for four developmental stages in triplicates: egg (W), second stage infective juvenile (J2), stage 3-4 juvenile (J3-J4) and adult female), *M. graminicola* (Dai et al. 2023) (data for four developmental stages in triplicates: W, J2, J3-J4 and adult female), *M. luci* (data for a mixture of stages with replicates in different conditions), *M. enterolobii* (Koutsovoulos et al. 2020) (data for two developmental stages in triplicates: J2 and J3-J4), *M. arenaria* (Blanc-Mathieu et al. 2017) (data for a mixture of stages without replicates), *M. javanica* (Blanc-Mathieu et al. 2017) (data for a mixture of stages without replicates). These data were aligned with the reference genome of each corresponding *Meloidogyne* species using STAR (Dobin et al. 2013). Then, RSEM (Li and

Dewey 2011) was used to determine expression values in FPKM (Fragments Per Kilobase Of Exon Per Million Fragments Mapped) for each gene. Thereafter, the results were normalized while maintaining positive expression values. The use of a base-10 logarithm allowed the data to be scaled without distorting the differences in value ranges or resulting in a loss of information. It is noteworthy that the addition of a value of 1 to the initial value yields positive expression values or zero, but never negative values.

When no replicates were available, the first quartile for expression values was calculated. Among the orphan genes in these species, those with expression values higher than this first quartile, calculated for the total gene set, were considered expressed. When replicates were available for specific stages or conditions, the median expression value for each gene was determined, also with the first quartile as a threshold to consider the gene as expressed. Finally, when a given gene was found to be expressed in at least one of the stages or conditions, it was considered to be expressed in the species under study.

5. Ribo-seq support for orphan genes

In addition to the transcriptomic evidence, we incorporated translatomic ribosome profiling (Ribo-seq) data from a recent study (Zhu et al. 2025) on *Meloidogyne incognita* to provide a first support for translation. The dataset included three developmental stages: pre-J2, J3–J4, and adult female, each with three biological replicates. Following the original authors' approach, we processed cleaned Ribo-seq reads: by aligning the reads to the *M. incognita* genome using STAR (Dobin et al. 2013) (--outFilterMismatchNmax 2, --outFilterMatchNmin 16, --alignEndsType EndToEnd) and performed transcript quantification with RSEM (Li and Dewey 2011), yielding FPKM values. Then, we treated each FPKM table of each developmental stage of each replicate in the exact way that we treated transcriptomics data. We normalized the tables and the first quartile of each FPKM value for each replicate is used as criteria to consider a gene as translated or not. When a given gene was found to be translated in at least one of the stages or conditions, it was considered to be translated. Finally, if there was at least one gene with evidence for translation in an orthogroup, all genes of the orthogroup were considered as translated.

6. Proteomic support for orphan proteins

To further support the existence of orphan proteins,, we used mass spectrometry-based proteomics dataset from the same study on *M. incognita* (Zhu et al. 2025). The data consists of different stages in three replicates: female, juvenile 2 and juvenile 3-4. We used the PEAKS studio software (Ma et al. 2003) to identify predicted proteins that have peptides matching with this mass spectrometry data for *M. incognita*. This calculation was applied to the predicted proteomes of *M. incognita* but also to *M. luci, M. javanica* and *M. arenaria* which are all closely related species from clade I. The other clade I species*, M. enterolobii,* was excluded because it is too distant from the others and does not share the same sub-genomes ancestry.

We considered a protein as detected if it showed a significantly matching peptide in at least one of these conditions. To ensure reliability, we filtered results based on statistical significance, retaining only those with p-values below 5% and FDR below 1%.

We then searched for orphan orthogroups that contained at least one protein with proteomic support. Similar to what was done with the transcriptomics and translatomics results, if a protein with mass spectrometry evidence belonged to an orphan orthogroup, we considered all proteins within that orthogroup to be supported by proteomics data.

7. Reconstructing the emergence dynamics of orphan genes

Two different tools, Bridge (Campos et al. 2024) and PastML (Ishikawa et al. 2019), were used to estimate the ancestral presence/absence status of each orthogroup along each branch of the species phylogenetic tree in order to investigate their emergence dynamics. PastML uses its own maximum likelihood (ML) method called MPPA, which accounts for branch lengths in the species tree. Bridge also implements its own algorithm but with gene loss more likely than gene gain. Both analyses were performed using the species tree inferred by OrthoFinder , which was based on homology relationships. To ensure the accuracy of our species tree, we compared it with a recently published mitochondrial tree for *Meloidogyne* species (Dai et al. 2023) and confirmed that the topologies were consistent. Bridge, implemented in R, was run with input data formatted as a table containing three columns (protein ID, species name, and orthogroup ID), along with a rooted species tree. The output consisted of a table listing orthogroups, nodes, and associated p-values. PastML was executed using its command line version, with input data structured as a matrix with orthogroups as columns and species as rows, along with a rooted species tree including branch lengths. The output was a tabular file containing node ids and presence of orthogroups at those nodes. The results from each tool were then filtered with a python script to identify the node of first occurrence of each orphan orthogroup on the branches of the species tree. Finally, the inferred number of emerging orthogroups at each node of the species tree was compared between the two methods.

Our dataset includes four species in Clade I that share recent common ancestry and parental A and B subgenomes (*M. incognita*, *M. luci*, *M. javanica* and *M. arenaria*). Our initial criterion of inclusion for orphan genes was presence in at least two *Meloidogyne* species. Therefore there is more chance for retention in this group than the rest of *Meloidogyne* species. To assess whether this could artificially inflate the number of emergence events in the branch leading to this group, we performed two additional analyses. In one analysis we kept only *M. incognita* as representative of this group and in another analysis, we kept only *M. incognita* and *M. luci*, the two most closely related species. Then, the emergence node of each orphan orthogroup is recalculated by Bridge in these two situations.

8. Identification of potential *de novo* genes specific to the *Meloidogyne* genus

8.1 Ancestral sequence reconstruction for each *Meloidogyne*-specific orthogroup

For each *Meloidogyne*-specific orthogroups protein sequences were aligned using MAFFT (Katoh 2002). A maximum likelihood phylogenetic tree was then constructed for each orthogroup using IQ-TREE (Minh et al. 2020), using the mset option to evaluate and select the best evolutionary model among WAG, JTT and LG. These three models were chosen because they are the only models available in both IQ-TREE and ARPIP (see below). Each tree was then rooted using TreeRecs (Comte et al. 2020) by reconciling it with the species tree. TreeRecs maps each gene in the gene tree to its corresponding species in the species tree and, by comparing the topology of the gene tree with that of the species tree, infers evolutionary events such as gene duplications and losses. TreeRecs then identifies the optimal root position in the gene tree that minimizes the number of inferred duplications and losses, ensuring a rooting consistent with the species tree. Finally, ancestral sequences for each orthogroup were reconstructed with ARPIP (Jowkar et al. 2023) using the best-fitting model identified by IQ-TREE. ARPIP was selected over alternatives like RaxML (Stamatakis 2014) and IQ-TREE because it treats gaps as separate character states, allowing meaningful inference of insertion and deletion events, whereas other tools treat gaps as missing data.

8.2 Alignment of ancestral protein sequences on the outgroup species

For orthogroups that contain at least a gene from clade III *Meloidogyne* species, the reconstructed ancestral protein sequences were aligned to the genome of *Pratylenchus penetrans*, the closest outgroup species with an annotated genome available using Exonerate (Slater and Birney 2005) (v 2.4.0). For other orthogroups shared by only a subset / Clade of *Meloidogyne* species, the ancestral proteins were aligned to the closest outgroup *Meloidogyne* species using the same tool. Exonerate splice aligns protein sequences to a genome translated in the 6 frames while accounting for possible introns. We used the default minimum score threshold of 100 and retained the best three alignments for each ancestral protein sequence.

To investigate the mechanisms of transition from a non-coding region to a transcribed coding gene, a home-made script was used to identify regions that aligned imperfectly according to Exonerate results. Specifically, we searched for premature stop codons, indels not in multiples of three, and incorrect splice sites in the outgroup genome, as these features disrupt correct transcription and translation.

Some genes detected in *Meloidogyne* may exist in the outgroup species but remain unidentified due to potential gene prediction errors. To account for this, all genes that aligned perfectly with the outgroup genome without any of the above problems were listed as potentially unpredicted genes in the outgroup. Using GFF3 annotation files for all genomes,

we verified whether these alignments corresponded to the position of a predicted gene in the outgroup genome. Using this approach, we classified our orphan genes as possible orphan genes that are highly diverged if a gene was already predicted at the same locus and had no homologs on the species that has the orphan gene, or *de novo* if no gene was predicted at the matching locus due to premature stop codon or other dysfunctional alignment.

8.3 Synteny analysis

We assessed whether orphan genes within the *Meloidogyne* genus were in a conserved synteny context using MCScanX (48). This step was essential to identify regions that are highly diverged or that may have given rise to *de novo* genes. This tool identifies collinear syntenic regions by comparing protein sequences using BLASTp results alongside corresponding genes positions via annotation files. MCScanX requires two concatenated input files, one containing all pairwise BLASTp results and another listing gene positions from annotation files. These files were generated using BLASTp results and GFF3 annotations corresponding to different comparisons and across all *Meloidogyne* species and the closest outgroup species, *P. penetrans*. Since most *Meloidogyne* genomes in Clade I are triploid or tetraploid, MCScanX was run with default parameters for simplicity, but with -b inter-species to identify collinear blocks only between different species, with an e-value of 0.001 and - s 4 to require at least four genes per collinear block. The tool produced a set of syntenic blocks along with a detailed list of syntenic gene pairs.

Following the initial analysis, the pairwise collinearity output file was reformatted to focus on syntenic blocks containing orphan genes. By default, MCScanX only reports collinear gene pairs, meaning genes without collinear homologs are absent from the output. For example, if a gene n and n+2 are collinear to y and y+1, we do not see the missing gene n+1 in the output file. To address this, we developed a home-made python script to identify genes missing from syntenic blocks. Using this, we pinpointed orphan genes on the collinearity file to be able to study the genes flanking the orphan genes. This analysis provided additional support for orphan gene classification by mapping both collinear gene positions and missing genes in the syntenic blocks.

Finally, we compared the list of candidate highly diverged orphan genes and the list of *de novo* gene candidates with synteny results using another home-made script. This allowed us to determine whether the flanking genes of the highly diverged/ *de novo* gene candidates were in conserved synteny between the species containing the gene and their outgroup.

9. Contribution of orphan genes to the set of parasitism effectors

To check whether orphan genes contribute to the set of parasitism effectors of *Meloidogyne* species, we retrieved the list of known effectors (a total of 111 effectors) from a previous study (Calia et al. 2024) where the effectors identified meet at least the criteria that the gene is

expressed in one of the secretion organs. Then we used cdhit2D (Li and Godzik 2006), with the parameters -s2 0 -c 0.90 -g1 -aL 0.30 -aS 0.30) in order to cluster known effector protein sequences with the predicted proteomes in our study. Then, by cross-referencing the list of effectors with the cd-hit clusters we defined which and how many orphan proteins correspond to effectors.

10. Characteristic features of orphan genes

Various properties of *Meloidogyne* orphan and non-orphan sequences were measured at both the gene and protein levels to enable comparisons and identification of features characteristic of orphans.

At the protein level, we measured:

(i) 22 physicochemical properties using a python script and the *Bio.SeqUtils.ProtParam* module from Biopython (27). These included the instability index, hydrophobicity profile, average flexibility, average hydropathy, isoelectric point, protein charge at pH 7.5, percentages of aliphatic amino acids and charged amino acids, molar extinction coefficient, cumulative percentages of small, aliphatic, aromatic, non-polar, polar, charged, basic and acidic amino acids in a helix, sheet or loop, and sequence length and molecular weight from protein sequences.

(ii) The abundance of each amino acid.

(iii) Three functional features: presence of InterPro protein domains detected with InterProScan (Zdobnov and Apweiler 2001), presence of secretory signal peptides detected with SignalP (Teufel et al. 2022) (v 6.0), and predicted subcellular localization with DeepLoc2 (Thumuluri et al. 2022).

At the gene level, we analyzed nucleotide frequencies and sequence lengths using a python script.

A summary table was created for each sequence by retrieving data from the output files of each script and tool (supplementary table 2). All these features were then compared between orphan and non-orphan sequences at both the protein and gene levels. Statistical tests were applied depending on the feature: Student's t-test or Mann-Whitney test for quantitative features (according to normality calculated with Shapiro test), and Fisher's exact test for qualitative features. P-values were included in the table for each feature, with values lower than 0.05 considered significant. Given the large dataset, even small differences might be significant. Therefore, for any feature that was significantly different, we calculated the relative as well as absolute differences in means between orphan and non orphan sequences for each feature to assess whether the difference can be considered different enough to have a biological relevance.

Furthermore, to understand the level of divergence among orphan and non-orphan proteins within orthogroups, sequence percent identities between *Meloidogyne* proteomes were

calculated using Diamond. The comparison is separated into two conditions: for orphan orthogroups and for non-orphan orthogroups. For the orphan orthogroups, we used the identified and expressed orphan orthogroups. The non-orphan orthogroups were selected using OrthoFinder and SonicParanoid results, where we considered the orthogroups with all *Meloidogyne* species present and at least 5 other nematode species. The following comparisons are then made between different species in the same orthogroups via a home-made script: proteins of Clade I species vs. each other, proteins of Clade II species vs. the other clades, proteins from Clade III species proteins vs. other clades, and for non-orphans we also compared proteins from all *Meloidogyne* clades vs. *P. penetrans* proteins. These percent ids are also taken into account to adjust parameters for ancestral sequence alignments.

Finally, the rate of non-synonymous (Ka) and synonymous (Ks) mutations between orphan and non orphan genes have been calculated using KaKs_Calculator (Zhang 2022) to compare their mutational retention patterns. As input, the tool requires CDS and protein sequences as well as pairs of homologs. As output the software returns a file with Ka and Ks values between pairs of homologs. We calculated these values for pairs of homologs of orphans and non orphans between *M. hapla* and *M. enterolobii*, *M. chitwoodi* and *M. hapla* to assess the levels of divergence when we transition from clade II to clade I and from clade III to clade II, respectively.

10. Classification of orphan genes and their origin using Random Forest

We trained two Random Forest classifiers (Breiman 2001) using the scikit-learn library (version 1.3.0) to assess whether orphan gene classifications could be predicted from their molecular characteristic features. We selected a subset of easily computable features (See data availability) from the full feature table. One classifier was built to distinguish between orphan and non-orphan sequences (as defined above), and the other was built to distinguish between highly diverged and *de novo* orphan sequences. For each task, we split the dataset into training and test sets and performed hyperparameter optimization using GridSearchCV with 5-fold cross-validation. Model performance was evaluated on the held-out test set by standard classification metrics precision, recall and F1-score. We extracted feature importances using the built-in feature_importances_ attribute of the trained classifiers to identify the variables that contributed most to classification.

Results

Almost 20% of *Meloidogyne* proteins are encoded by transcriptionally supported orphan genes

Nearly 25% of the predicted *Meloidogyne* proteins do not have identifiable homologs in other nematodes

Since most nematode species' genomes are not present in generalist public databases, we first identified *Meloidogyne*-specific proteins, in comparison to other nematodes. From the available proteomes, we selected 85 representative nematode species using a quality criterion based on a minimum metazoa BUSCO score of 62% and the absence of contaminants according to OMArk, as well as a species representativeness criterion, covering a total of 44 genera and 10 clades (supplementary table 1). Of the nine *Meloidogyne* species for which a proteome was available, eight were retained, covering the three main *Meloidogyne* clades (I, II and III). After an all-aginst-all comparison, the best reciprocal homologs were then classified into orthogroups by OrthoFinder and SonicParanoid. OrthoFinder classified 92% of the 1,970,290 proteins from the 85 nematodes into a total of 275,056 different orthogroups, while SonicParanoid classified 79% of all these proteins into 75,469 different orthogroups. We selected orthogroups that exclusively contained *Meloidogyne* proteins and eliminated orthogroups that were specific to a single species to minimize the risk of prediction errors in proteomes. This led to the identification of 15,024 and 14,955 orphan orthogroups according to OrthoFinder and SonicParanoid, respectively. These orthogroups contain a total of 81,914 and 64,342 orphan proteins, corresponding to 26% and 21% of the proteins predicted in *Meloidogyne* genomes. These *Meloidogyne*-specific proteins are possibly encoded by orphan genes with no homologs in nematodes other than *Meloidogyne*.

Most orphan genes are supported by expression data

Although presence of homologs in at least two different *Meloidogyne* species minimizes the risk of prediction errors, it is still possible that the same error has been made by prediction software in two species. To further minimize such prediction errors, we only retained orphan proteins that were encoded by genes with substantial expression support from RNA-seq transcriptome data.

From the list of orphan orthogroups, all *Meloidogyne*-specific orthogroups were analyzed accordingly, and only those for which at least one gene was supported by transcriptomic data were retained. The analysis identified a total of 13,279 and 13,185 orthogroups (respectively for OrthoFinder and SonicParanoid) containing at least one orphan protein whose corresponding gene was supported by transcriptomic data. This represents a total of 76,930 and 59,859 orphan proteins supported by expression data, respectively, encompassing 25% and 19% of all proteins predicted for the genus *Meloidogyne*.

Most *Meloidogyne* orphan proteins relative to other nematodes also lack homologs in the rest of living organisms

The next verification step was to determine whether, beyond nematodes, *Meloidogyne*-specific proteins are orphans among the rest of living organisms. A homology search across the domains of life represented in the NCBI nr protein database, was performed to rule out two cases: (i) genes possibly acquired by horizontal transfer events between *Meloidogyne* and species other than nematodes, and (ii) genes present in an older common ancestor but lost in all nematodes except *Meloidogyne,* although this scenario seems less likely. Most of the proteins predicted to be orphans in *Meloidogyne* nematodes either had no hits with any proteins in nr or had hits only with proteins from *Meloidogyne* species, which is not surprising considering that nr contains the proteome of only a few *Meloidogyne* species. Only 273 proteins were found to have a significant hit outside the *Meloidogyne* species and were classified as non-orphans for further analysis. We then eliminated all orthogroups obtained from OrthoFinder and SonicParanoid that contained at least one of these non-orphan proteins, resulting in a total of 76,717 and 59,488 orphan proteins, corresponding to 24% and 19% of all *Meloidogyne* proteins, respectively.

16% *Meloidogyne* proteins are transcriptionally verified orphans

To ensure consistency in using both OrthoFinder and SonicParanoid to identify orphan proteins, we compared the results from both software after verifying transcription and confirming the absence of homology with other organisms. Instead of directly intersecting orphan protein sets, we focused on orphan orthogroups, which are more informative for downstream analyses. We compared all orphan orthogroups identified by both tools and retained those that were either identical, that shared at least 75% of proteins, or were subsets of each other based on sequence composition. This approach left us with 8,974 non-redundant and transcriptionally verified orphan orthogroups, comprising a total of 48,681 orphan proteins, which represent 16% of the *Meloidogyne* proteome (Figure 2). Depending on the species, this percentage ranges between 6% to 21% with the higher proportions in the parthenogenetic polyploid species in clade I.

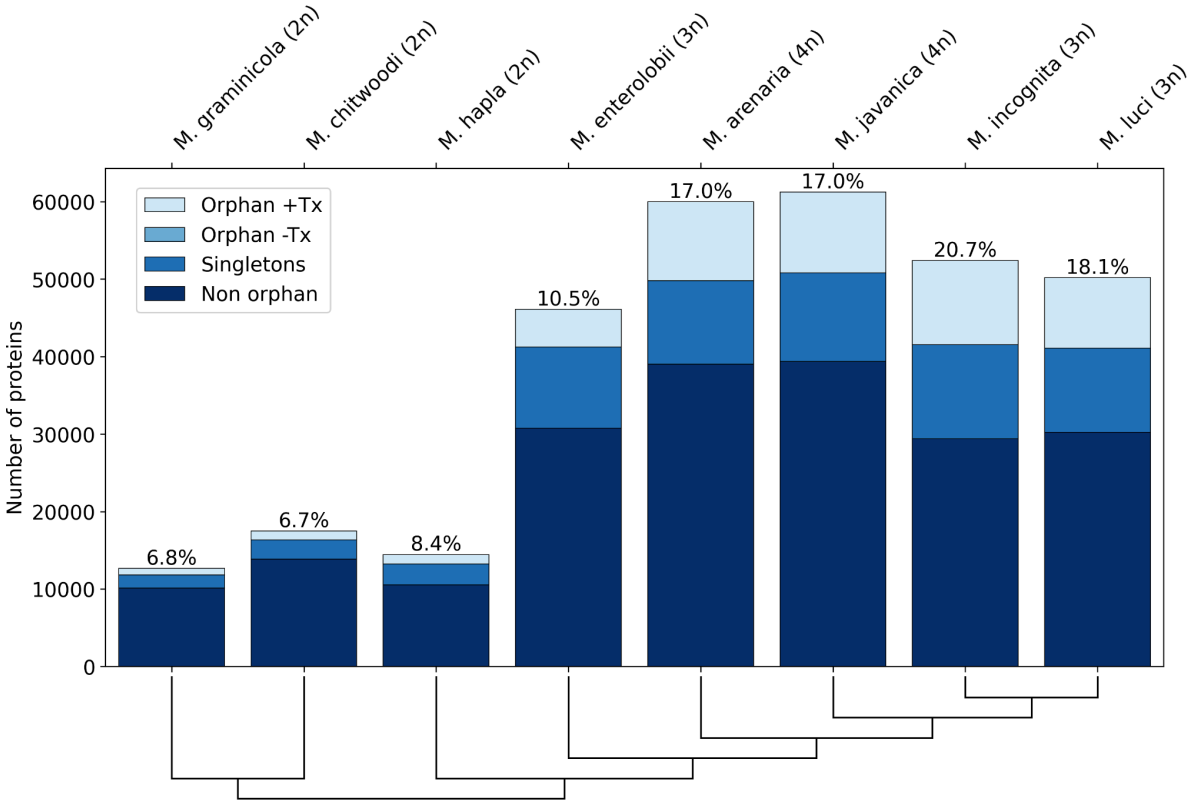


Figure 2: Number of orphan genes and their origin for 8 *Meloidogyne* species studied. Barplot representing the number of orphan genes for each of the eight *Meloidogyne* species. -Tx/+Tx represents transcriptomic validation. However Orphan -Tx cases are not visible on the graph as there are very few. The percentages on top of the bars for each species correspond to the fraction of orphan proteins with transcriptional evidence. A species tree with clade separation is shown below the plot.

Some orphan proteins are further supported by Ribo-seq and proteomics data

Once the transcriptionally-supported *Meloidogyne* orphan genes were identified, our next objective was to identify evidence for translation. To do so, we used ribo-seq and proteomics data from a recent study on *M. incognita* (Zhu et al. 2025). The ribo-seq data provided evidence for translation for 31,595 orphan genes. Besides ribo-seq, a total of 17,596 proteins from Clade I *Meloidogyne* out of the 200,000 predicted proteins, were supported by proteomic data, regardless of their orphan status. Then, filtering for the orphans and assuming that proteomic support within an orphan orthogroup indicates that the entire orthogroup is supported, we identified 1134 orphan proteins whose existence is supported by proteomic data. These numbers are not much neither for orphans nor for non-orphans, considering the total number of orphan genes identified in clade I and the total size of clade I genomes because of the fact that we have mass spectrometry data only for *M. incognita* and not for the other clade I species. Therefore, it is clear that we are missing translational evidence at the protein level but at least this analysis showed that we have real evidence of some orphan

proteins being translated. When we took the union of supported translation either from ribo-seq or proteomics, we have in total 31,607 orphan proteins with translational evidence. A venn diagram summarizing transcriptional and translational evidence can be found in supplementary (supplementary figure 2).

The emergence of orphan genes in *Meloidogyne* shows several phases of acceleration

To understand the dynamics of orphan gene emergence in *Meloidogyne*, we estimated the most likely time of appearance for each transcriptionally-supported orphan orthogroup on the species tree, while accounting for potential gene losses, in some species using two different methods, Bridge and PastML (Figure 3).

Despite some differences in the number of orthogroups predicted to have emerged at each node—particularly within Clade I (*M. javanica, M. luci,* and *M. incognita*)—the overall trend was highly similar between both methods. Our analysis revealed a steady increase in the number of orphan orthogroup emergence from the common ancestor to individual species. However, this increase was not uniform, as we identified several bursts of emergence, notably in the ancestor of Clade I and, more prominently, in the common ancestor of *M. arenaria, M. javanica, M. luci,* and *M. incognita*, where both methods predicted the highest number of emerged orphan, orthogroups (Figure 3).

It should be noted that these three species share common ancestral A and B subgenomes that they inherited from recent hybridization events (Blanc-Mathieu et al. 2017; Dai et al. 2023; Mota et al. 2024). Therefore, if an orphan gene emerged ancestrally in A or B, it will likely be inherited by all the descendent species. Since one criterion for orphan gene retention in our analysis was presence in at least two different species, this has more chance to happen in this group of closely related species.*M. enterolobii*, on the other hand, has not been described so far as sharing the same A and B subgenomes, and according to mitochondrial-based phylogeny is quite distant from the rest of Clade I species (Blanc-Mathieu et al. 2017).

This pattern aligns with observations from the species tree inferred by OrthoFinder (supplementary figure 3), where *M. enterolobii* appears more distantly related to other Clade I species based on branch lengths. Additionally, when analyzing the percent identities of orphan proteins within orthogroups, we found that orphan proteins from Clade I species (excluding *M. enterolobii*) shared a mean identity of 91.2% when compared among each other. However, when compared to *M. enterolobii*, this percentage dropped to 72%, confirming that *M. enterolobii* is substantially more divergent than the rest of Clade I species.

As presence of four closely related Clade I species might artificially inflate the number of orphan gene emergence compared to Clade II and III, we reproduced the same analyses keeping only two or only one representative from these four species (supplementary figure 1). In both cases, the ancestral state reconstruction consistently confirmed a burst of orphan gene

emergence in the branch leading to Clade I. These findings support the hypothesis of a burst of more than 1000 orphan gene families emerging in an ancestor of Clade I.

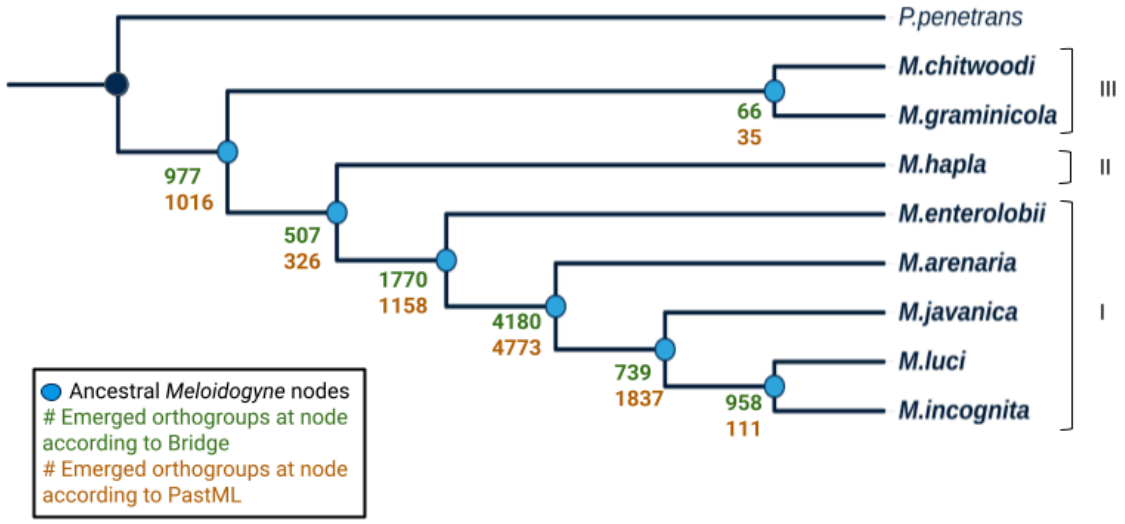


Figure 3: Number of emerged orphan orthogroups at each node of the species tree of *Meloidogyne* species with their closest outgroup species *P.penetrans.* Species tree with 8 *Meloidogyne* species corresponding to three different clades (represented at the right of the species) of *Meloidogyne* genus along with *P. penetrans*, which is the closest outgroup to this genus. Each ancestral *Meloidogyne* node is represented by a blue dot. Number of orphan orthogroups emerged in each node is given by a value next to it, in green for Bridge and in orange for PastML.

De novo gene birth accounts for almost 20% of orphan genes in *Meloidogyne*

To identify the origin of orphan genes and tell apart *de novo* gene birth from high divergence from a pre-existing gene, comparison to an outgroup species and synteny analysis are essential steps. In order to improve the chance of aligning orphan proteins to outgroup genomes that do not contain that orphan protein, we reconstructed ancestral protein sequences for each orphan orthogroup. Doing so, the reconstructed ancestral sequence is in principle phylogenetically closer to the outgroup. The *Pratylenchus* genus is the closest to the *Meloidogyne* and *P. penetrans* the closest species with a genome sequence available. Therefore, for all orphan orthogroup that were deemed present in the last common ancestor of the *Meloidogyne* species (Figure 3), we splice-aligned the reconstructed ancestral protein to the genome of *P. penetrans*, translated in the 6 frames. For orphan groups that were not at the root, but specific to a *Meloidogyne* clade or branch, we splice-aligned the reconstructed protein to the closest outgroup *Meloidogyne* species genome.

To identify genes that may have emerged *de novo*, we examined genomic regions in the outgroups that matched with the ancestral sequence but showed features incompatible with

gene splicing or translation, such as premature STOP codons, indels not in multiples of three, or incorrect splice signals. It should be noted here that, by definition, beyond the closest outgroup, the orphan gene is absent from all the other nematodes and even the rest of species. Mutations in these non-genic regions could have allowed the progressive emergence of a coding gene. We further detail here an example of an orphan gene common to all Clade I *Meloidogyne* species except *M. enterolobii.* In this case, the inferred ancestral orphan protein was aligned on the *M. enterolobii* genome, where we observed a whole set of incompatible features for gene translation on this outgroup. Mutations arose in an ancestor of the other species in Clade I allowing a sequence compatible with transcription and translation (figure 4a). Obviously, the gene is also absent from *M. hapla* (Clade II) as well as *M. graminicola* and *M. chitwoodi* (Clade III).

Using gene annotation files for all genomes we checked if the inferred ancestral orphan proteins were aligning to genomic regions compatible or not with the presence of a canonical gene in the outgroups. Firstly, it is important to note that around 10% of our ancestral sequences were aligning on the outgroup almost perfectly, and compatible with correct transcription and translation. Therefore, these ones were considered false positive orphans as these genes may simply not have been predicted in the closest outgroup species. Using conserved synteny analysis with the flanking genes; we found that, for 20% of the orphan genes, another gene was present at the expected locus of the orphan gene in the outgroup (figure 4b). Therefore, we can hypothesize that they have strongly diverged from a common ancestral gene, impairing identification of homologs. On the other hand, we identified at least 2,031 ancestral orphan proteins that aligned on outgroup species but in a way not compatible with a functional gene. The 8,833 corresponding genes likely emerged *de novo* in *Meloidogyne* species at different evolutionary time points, which corresponds to 18% of *Meloidogyne* orphan genes. These observations suggest that *de novo* gene birth is responsible for an important part of the presence of orphan genes in the *Meloidogyne* genus. The rest of the ancestral orphan sequences could not be aligned to the closest outgroup species. Therefore, for these cases, we cannot tell if they are highly diverged orphans or *de novo* ones (figure 5).

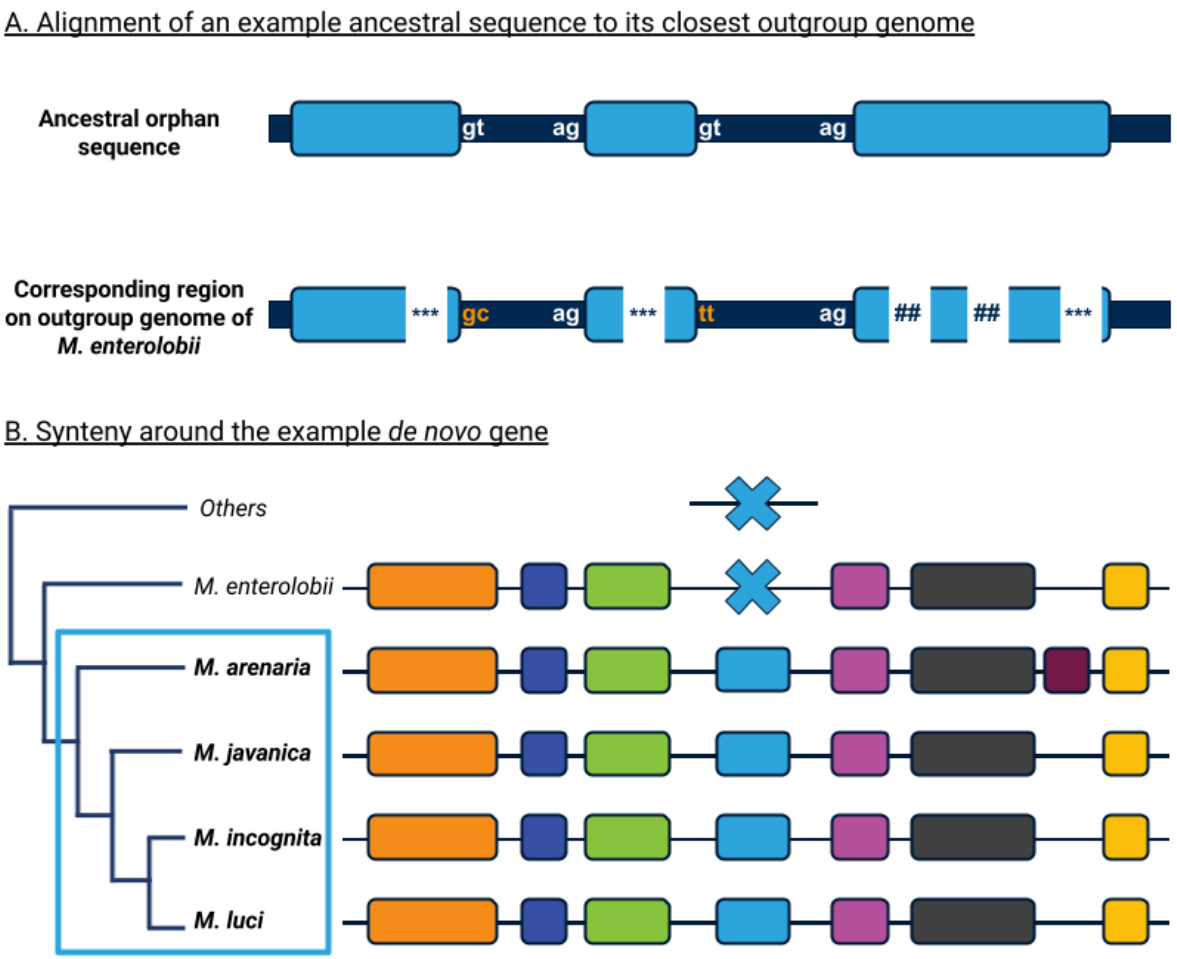


Figure 4: Example of a *de novo* gene birth that emerged inside clade I. Alignment of the reconstructed ancestral sequence of an example orthogroup common to all members of clade I except *M. enterolobii* to the genome of *M.enterolobii* using Exonerate. Dark blue represents introns and light blue represents exons. Stop codons are represented by ***. Frameshift insertions or deletions are re[represented by #. Correct splicing sites are in white, mutated ones in orange. B. Verification of the syntenic conservation around the identified *de novo* gene. For the position where ancestral sequence aligns to *M.enterolobii*, the closest 3 upstream and downstream genes are studied to determine the conservation of the region. Lines represent intergenic regions and boxes correspond to the genes. Light blue boxes correspond to the *de novo* gene and blue cross is the locus it aligns with on *M.enterolobii*. The color code indicates where the same genes align on different species. The blue rectangle on the species tree represents the species that contain the *de novo* gene.

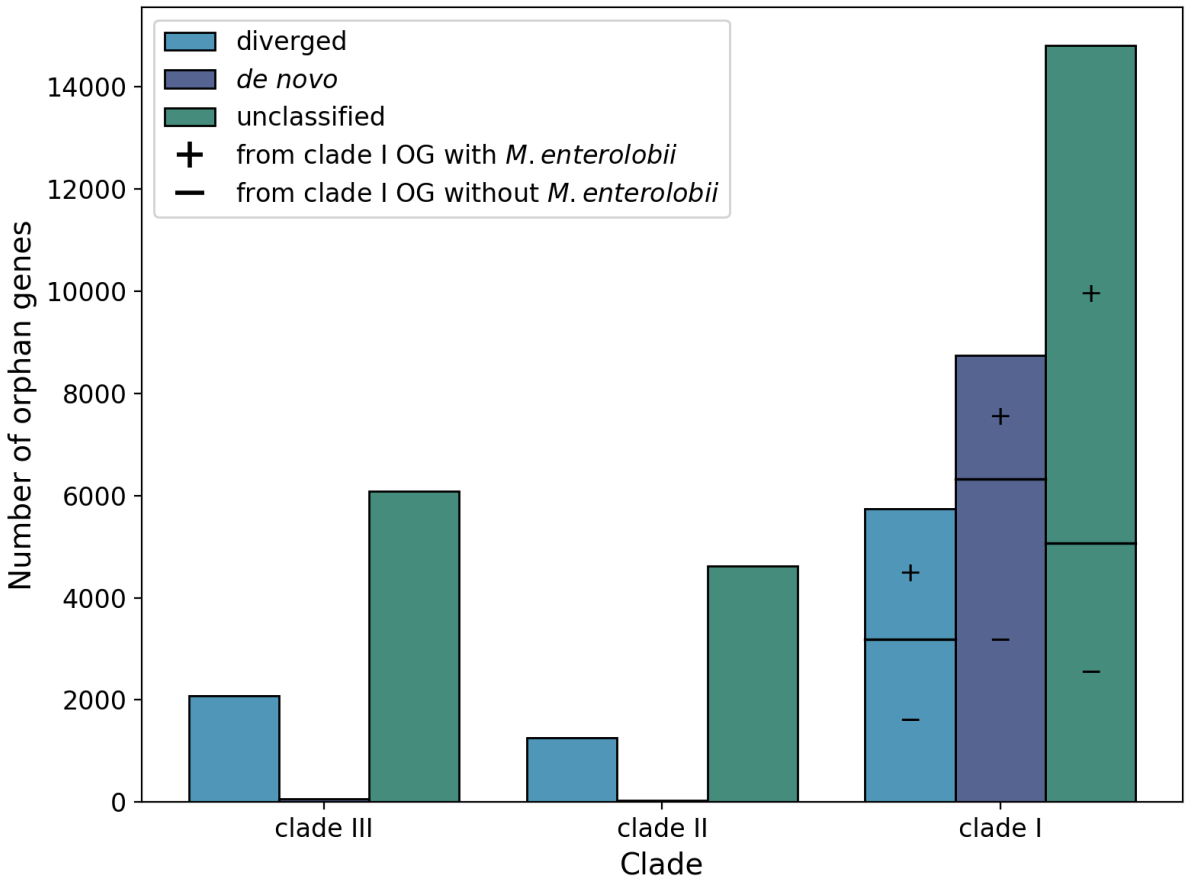


Figure 5: Number of orphan genes identified as highly diverged, *de novo* or unclassified. The number of orphan genes for each clade and classification is represented by barplots. For Clade I species, there is also a distinction between orthogroups (OGs) that contain *M. enterolobii* and those that do not.

Orphan genes contribute to the parasitism of root-knot nematodes

By comparing *Meloidogyne* predicted proteins to a reference dataset of 111 known effectors from a previous study (Calia et al. 2024), we recovered 83 clearly matching effectors under strict clustering parameters. Among these, 34 were encoded by orphan genes (including 9 singletons) representing approximately 40% of all matched effectors. This suggests that nearly half of the known parasitism-related effectors in *Meloidogyne* are of orphan origin. This correlates well with previous reports indicating that the majority of effectors secreted by nematodes are encoded by orphan genes (Mitchum et al. 2013a). Therefore, *de novo* gene birth and high divergence from pre-existing genes seem to have substantially contributed to the emergence of the parasitic arsenal of these root-knot nematodes.

Orphan genes tend to encode short secreted proteins

By measuring physicochemical, compositional, and functional features for orphan and non-orphan *Meloidogyne* genes and their corresponding protein sequences, we identified features that differ between orphans and non-orphans (supplementary table 2). Given the large data set, nearly all measured features showed statistically significant differences. However, some of these differences were too small to be biologically meaningful. To refine our analysis, we filtered the results based on relative and absolute mean differences between orphans and non-orphans to see if these differences might have a biological importance or not.

Orphan protein sequences were found to be generally shorter, exhibited differences in amino acid composition, especially in lysine and arginine where orphan proteins have more lysine and less arginine in their proteins, and were more positively charged at pH 7.5 compared to non-orphans. Additionally, orphan proteins were more frequently predicted to be extracellular or localized in plastids. This correlates well with the results of a previous study that demonstrated that a high lysine to arginine ratio is associated with extracellular localization (Warwicker et al. 2014).

Orphan proteins contained fewer known domains, which is expected due to their lack of homologs. A notable difference concerned the prediction of signal peptides: 4.1% of the orphans were predicted to contain a signal peptide compared to only 2.4% for non-orphan proteins. In addition, orphan proteins accounted for almost 30% of proteins with a predicted signal peptide, indicating an enrichment of signal peptides among orphan proteins. This aligns with our observation that orphan proteins were more frequently predicted to be extracellular and correlates well with our observation that nearly half of known effectors secreted by root-knot nematodes are encoded by orphan genes (Mitchum et al. 2013b).

We also measured Ka and Ks values between pairs of homologous genes in pairwise comparisons of Clade III vs. Clade II *Meloidogyne* and Clade II vs. Clade I. We did this comparison both for orphans and non orphan genes. For orphans, we had an average Ka of 0.39, Ks of 2.79 and Ka/Ks of 0.19 for the comparison between clade III and clade II, and we had a Ka of 0.32, Ks of 1.62 and Ka/Ks of 0.31 for the comparison between clade II and clade I. In comparison, for non orphans, the mean values were respectively Ka 0.12, Ks 1.92, Ka/Ks 0.07 and Ka 0.21, Ks 1.52, Ka/Ks 0.12. The results showed that for each situation, Ka, Ks and Ka/Ks values are higher for orphans compared to non orphans. Hence, orphan genes tend to accumulate more mutations, of which a higher proportion are non-synonymous, suggesting relaxed selection on these relatively recent genes.

According to all these measurements and statistics, we showed that orphan genes and proteins in *Meloidogyne* differ from non-orphans. Most importantly, orphan proteins are shorter, more positively charged, have an enrichment of signal peptides and have a high K/R ratio which explains why they tend to be more extracellular (Warwicker et al. 2014). In other species, different features between orphans and non-orphans were also observed, especially the

difference in sequence length and GC content (which was not the case for us) but here we also demonstrated some additional features for the case of orphan proteins in root-knot nematodes (Fakhar et al. 2023).

Random Forest classifiers discriminate orphan/non-orphan and *de novo*/diverged proteins

Given the computational cost of our full comparative pipeline and the observation that orphan genes display distinct molecular characteristics, we sought to evaluate whether these features could be used to rapidly classify orphan genes using machine learning. We therefore trained two Random Forest classifiers: one to distinguish orphan vs. non-orphan genes, and another one to distinguish highly diverged vs. *de novo* orphan genes.

The orphan vs. non-orphan classifier achieved an F1 score of 0.90, with a precision of 0.87 and recall of 0.93, indicating high overall performance with a slightly stronger ability to retrieve true positives (recall) than to avoid false positives (precision). The *de novo* vs. highly diverged classifier also performed well, with an F1 score, precision, and recall all at 0.90, suggesting a well-balanced and robust model for separating these two categories of orphan genes.

Feature importance analysis revealed that the most influential variables for classifying orphans versus non-orphans were sequence length, probability of being extracellular, and lysine frequency. For the *de novo* vs. highly diverged classifier, the most impactful features were probability of being localized at mitochondrion, frequency of acidic amino acids, sequence length, and phenylalanine frequency. While these features were the top contributors, a wide range of other features also influenced the classification decisions in both models, underscoring the multifactorial nature of orphan gene characteristics.

These results demonstrated that we have successfully built classifiers to rapidly identify orphan genes and also their origin (*de novo* vs. highly diverged) on root-knot nematodes. This can be extremely useful as these classifiers are extremely fast compared to our comparative genomics pipeline and unlock the possibility to scan massive datasets of predicted proteomes. Furthermore, we also applied the highly diverged/*de novo* classifier to the cases where we couldn't determine the origin and were initially left as unclassified with our pipeline. This was tested for each clade: For clade I, 56% out of 14807 unclassified orphans were classified as *de novo*, whereas for clade II only 10% out of 4617 unclassified ones were predicted as *de novo*. Finally, for clade III and universally conserved orphans across *Meloidogyne*, only 8% out of 6082 unclassified ones were predicted as *de novo*. These results suggest a higher prevalence of *de novo* genes in clade 1, which is consistent with the manual classification (Figure 5). In addition, this classification demonstrated that 2 of the known effectors are probably encoded by *de novo* genes.

Discussion

Since 2008, orphan genes in root-knot nematodes have attracted the attention of researchers. This is for two main reasons (i) being specific to *Meloidogyne* species; these genes are possibly involved in plant parasitism, and (ii), because these genes have no homologs in the other species, specifically targeting them has low chance to have undesired side effects in the other species (Danchin et al. 2013b; Grynberg et al. 2020b). However, despite their high interest, their origin, evolutionary history and characteristics had remained mysterious and unexplored so far.

In this study, we characterized orphan genes in *Meloidogyne* species, in the light of more recent genome data and comparative genomics software. More importantly, we traced back for the first time their evolutionary origins and trajectories as well as their characteristic signatures. We found that approximately 16% of the genes in *Meloidogyne* species are transcriptionally supported, genus-specific orphan genes. This proportion is consistent with the 5-30% range previously reported for eukaryotic genomes (Tautz and Domazet-Lošo 2011). Direct comparison of these percentages between different studies done at different times with different methodologies remains difficult. However the substantially high fraction we observed suggests a functional and evolutionary importance of orphan genes in root-knot nematodes. As a conservative strategy to avoid over-predictions of gene annotation software, we have discarded species-specific orphan genes and only kept those present in at least two species. Therefore, the total number of orphan genes in these root-knot nematodes is certainly higher. Furthermore, it would be interesting in the future to confirm whether some of these genes are actually species-specific as they might represent a good potential for the development of markers to differentiate the *Meloidogyne* species. Differentiating these species based on molecular and morphological markers is challenging yet important for the farmers to adapt crop rotation and pest management strategies.

Our comparative analyses were performed using 85 quality-checked nematode proteomes, representing a comprehensive and diversified dataset. However, it should be kept in mind that the number of orphan genes identified remains sensitive to the completeness and diversity of available genomes. As new high-quality genomes become available, especially from lineages closely related to *Meloidogyne*, our ability to identify orphan genes as well as their origin will improve. The *Pratylechus* genus is the most closely related relative to the *Meloidogyne* genus (Holterman et al. 2017; Ahmed et al. 2022; Qing et al. 2024), and by the time we performed this comparative analysis, *Pratylenchus penetrans* was the only species with an available annotated genome passing our completeness threshold (Mota et al. 2024). Even though genome papers have been published for three other *Pratylenchus* species, *P. coffeae* (Burke et al. 2015), *P. scriberni* (Arora et al. 2023) and *P. vulnus* (Dai et al. 2025) , the annotation data are not publicly available and could not be used. Inclusion of additional *Pratylenchus* species may reveal closer relatives to *Meloidogyne* and refine our inferences regarding gene novelty. This limitation is particularly important for our synteny-based approach, which is currently more effective when applied within the genus *Meloidogyne*. Indeed, when we apply synteny within *Meloidogyne* species to study *de novo* gene birth within the genus, we have more

confident alignments and synteny results. However, when we investigate the origin of orphan genes present in the last common ancestor of *Meloidogyne* species, it is more challenging to establish a synteny around orphan genes of *Meloidogyne* with *P. penetrans* genome. This is due to the higher evolutionary distance between *Pratylenchus* and the *Meloidogyne* than between the three clades within the *Meloidogyne*. Indeed, the last common ancestor of *Meloidogyne* Clades I, II and III is estimated to be ca. 65 My old, while the last common ancestor of these *Meloidogyne* and *P. penetrans* is estimated at ca. 136 My ago (Qing et al. 2025). This is also reflected by lower percent identities between orthologous proteins in *Pratylenchus* vs. *Meloidogyne* as compared to within *Meloidogyne* identities. Overall, this highlights the need for a better genomic representation of intermediate lineages to resolve the deeper evolutionary origins of these genes.

Using mass spectrometry and translatomic data, we were also able to verify the translation of more than half of transcriptionally-supported orphan genes, providing further evidence that these genes are not only transcribed but also translated as proteins. However, the availability of scarce proteomics data and for only one *Meloidogyne* species limits our ability to validate more broadly the existence of these proteins across the genus. Also, even for the non-orphans, it is clear that we are missing translational evidence. Expanding proteomic coverage would greatly enhance future studies.

In terms of gene origin, we estimated that high divergence from homologous genes, and *de novo* gene birth,, respectively account for 20% and 18% of orphan genes. Nevertheless, for roughly half of the orphan genes, the mechanism of origin remains ambiguous. For these cases, it is difficult to determine which mechanism is involved for two main reasons: (i) we cannot always successfully align the orphan sequence to the outgroup genome and (ii) the synteny is not always conserved around the orphan gene. Moreover, it is important to note that gene emergence is not always attributable to a single mechanism; both high divergence and *de novo* processes may jointly contribute to the presence of an orphan gene (Prabh and Rödelsperger 2019). However, when we applied our Random Forest classifier to these ambiguous cases, we observed that in clade I, *de novo* gene birth is more prevalent whereas for other clades having high divergence from ancestral genes is more prevalent. This might happen because of the fact that clade I species are more recent or whole-genome duplication in these species may be favoring *de novo* gene birth. Polyploidy in these species is due to complex hybridization events (Blanc-Mathieu et al. 2017; Szitenberg et al. 2017; Dai et al. 2023) and these events can be associated with re-activation and multiplication of transposable elements. Recent findings have confirmed that transposons can favor de novo transcripts emergence and eventually de novo gene birth (Lebherz et al. 2024). Another possibility is that as most of the identified *de novo* genes that we had identified by synteny analysis are in clade I, the classifier was trained to identify them and therefore it may have difficulties to classify *de novo* genes in other clades.

Our results also show that orphan genes as a whole encode proteins with distinctive features compared to non-orphans. Specifically, orphan proteins tend to be shorter, more positively charged, enriched in signal peptides, and have increased lysine/arginine (K/R) content. These

characteristics are consistent with proteins that function in the extracellular environment and support the observation that many effectors secreted by plant-parasitic nematodes are encoded by orphan genes. This also suggests other effectors remain to be discovered among the dataset of *Meloidogyne*-specific orphan genes. While this hypothesis is compelling, further experimental validation is needed to determine the extent to which orphan genes contribute to the set of effector proteins and more generally to parasitism in these nematodes.

Overall, our study provides the first comprehensive analysis of orphan genes, including investigation of their origin (high divergence or *de novo* gene emergence) and their characteristic features in *Meloidogyne*, a genus of agriculturally important plant-parasitic nematodes. We show that orphan genes comprise an important fraction of *Meloidogyne* gene sets, with a substantial proportion likely to have emerged *de novo*. These findings advance our understanding of how new genes arise and evolve in parasitic lineages and highlight the importance of orphan genes in the adaptive biology of *Meloidogyne*. Future studies should focus on the functional validation of candidate effectors among orphan genes, and the expansion of phylogenomic and proteomic datasets to refine evolutionary inferences.

Code Availability

The scripts written to treat the outputs of different tools as well as to treat and filter different results can be found on GitHub upon submission of the manuscript (repository: erseckin/melo-evo-orphan).

Data Availability

All datasets generated in this study are deposited in the « Orphan and *De Novo* Genes in Root-Knot Nematodes » dataverse on recherche.data.gouv.fr and have been assigned persistent DOIs; they will be made publicly available upon submission of the manuscript.

Acknowledgements

This research was supported by the joint INRAE-Inria PhD program, which funds the PhD thesis of Er.S. We are grateful to the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, DOI:10.15454/1.5572369328961167E12) as well as to the bioinformatics and genomics platform, BIG, Sophia Antipolis (ISC PlantBIOs, DOI: 10.15454/qyey-ar89) for computing and storage resources. We would also like to thank Gianni Liti and Mathilde Carpentier for their suggestions and ideas that were discussed during Er.S.'s Ph.D. follow-up committee meetings.

References

Abad P, Gouzy J, Aury J-M, Castagnone-Sereno P, Danchin EGJ, Deleury E,

Perfus-Barbeoch L, Anthouard V, Artiguenave F, Blok VC, et al. 2008. Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. *Nat. Biotechnol.* 26:909–915.

Ahmed M, Roberts NG, Adediran F, Smythe AB, Kocot KM, Holovachov O. 2022. Phylogenomic Analysis of the Phylum Nematoda: Conflicts and Congruences With Morphology, 18S rRNA, and Mitogenomes. *Front. Ecol. Evol.* 9:769565.

Arora D, Hernandez AG, Walden KKO, Fields CJ, Yan G. 2023. First Draft Genome Assembly of Root-Lesion Nematode Pratylenchus scribneri Generated Using Long-Read Sequencing. *Int. J. Mol. Sci.* 24:7311.

Blanc-Mathieu R, Perfus-Barbeoch L, Aury J-M, Da Rocha M, Gouzy J, Sallet E, Martin-Jimenez C, Bailly-Bechet M, Castagnone-Sereno P, Flot J-F, et al. 2017. Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes.Gojobori T, editor. *PLOS Genet.* 13:e1006777.

Breiman L. 2001. Random Forests. *Mach. Learn.* 45:5–32.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12:59–60.

Burke M, Scholl EH, Bird DMcK, Schaff JE, Colman SD, Crowell R, Diener S, Gordon O, Graham S, Wang X, et al. 2015. The plant parasite Pratylenchus coffeae carries a minimal nematode genome. *Nematology* 17:621–637.

Cai J, Zhao R, Jiang H, Wang W. 2008. *De Novo* Origination of a New Protein-Coding Gene in *Saccharomyces cerevisiae*. *Genetics* 179:487–496.

Calia G, Porracciolo P, Chen Y, Kozlowski D, Schuler H, Cestaro A, Quentin M, Favery B, Danchin EGJ, Bottini S. 2024. Identification and characterization of specific motifs in effector proteins of plant parasites using MOnSTER. *Commun. Biol.* 7:850.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Campos LRS, Trefflich S, Morais DAA, Imparato DO, Chagas VS, Albanus RD, Dalmolin RJS, Castro MAA. 2024. *Bridge* : A New Algorithm for Rooting Orthologous Genes in Large-Scale Evolutionary Analyses.Bonatto S, editor. *Mol. Biol. Evol.* 41:msae019.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423.

Comte N, Morel B, Hasić D, Guéguen L, Boussau B, Daubin V, Penel S, Scornavacca C, Gouy M, Stamatakis A, et al. 2020. Treerecs: an integrated phylogenetic tool, from sequences to reconciliations.Ponty Y, editor. *Bioinformatics* 36:4822–4824.

Cosentino S, Iwasaki W. 2023. SonicParanoid2: fast, accurate, and comprehensive orthology inference with machine learning and language models. Bioinformatics Available from: http://biorxiv.org/lookup/doi/10.1101/2023.05.14.540736

Dai D, Xie C, Zhou Y, Bo D, Zhang S, Mao S, Liao Y, Cui S, Zhu Z, Wang X, et al. 2023. Unzipped chromosome-level genomes reveal allopolyploid nematode origin pattern as unreduced gamete hybridization. *Nat. Commun.* 14:7156.

Dai D, Zhang Y, Latina R, Yang X, Williamson VM, Groen SC, Shamsullah S, Leslie CA, Castro B, Siddique S. 2025. Genomic and Transcriptomic Insights into the Evolution and Parasitic Strategy of the Woody-Plant Nematode *Pratylenchus vulnus*. Available from: http://biorxiv.org/lookup/doi/10.1101/2025.09.24.678327

Danchin EGJ, Arguel M-J, Campan-Fournier A, Perfus-Barbeoch L, Magliano M, Rosso M-N, Da Rocha M, Da Silva C, Nottet N, Labadie K, et al. 2013a. Identification of Novel Target Genes for Safer and More Specific Control of Root-Knot Nematodes from a Pan-Genome Mining.Ghedin E, editor. *PLoS Pathog.* 9:e1003745.

De Ley IT, De Ley P, Vierstraete A, Karssen G, Moens M, Vanfleteren J. 2002. Phylogenetic Analyses of Meloidogyne Small Subunit rDNA. *J. Nematol.* 34:319–327.

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.

Dujon B. 1996. The yeast genome project: what did we learn? *Trends Genet.* 12:263–270.

Eisenback JD, Triantaphyllou HH. 2020. Root-Knot Nematodes: Meloidogyne Species and Races. In: Nickle WR, editor. Manual of Agricultural Nematology. 1st ed. CRC Press. p. 191–274. Available from: https://www.taylorfrancis.com/books/9781000104509/chapters/10.1201/9781003066576-6

Ekstrom A, Yin Y. 2016. ORFanFinder: automated identification of taxonomically restricted orphan genes. *Bioinformatics* 32:2053–2055.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20:238.

Fakhar AZ, Liu J, Pajerowska-Mukhtar KM, Mukhtar MS. 2023. The Lost and Found: Unraveling the Functions of Orphan Genes. *J. Dev. Biol.* 11:27.

Grandchamp A, Aubel M, Eicholt LA, Roginski P, Luria V, Karger A, Dohmen E. 2025. *De Novo* Gene Emergence: Summary, Classification, and Challenges of Current Methods.Vakirlis N, editor. *Genome Biol. Evol.* 17:evaf197.

Grynberg P, Coiti Togawa R, Dias De Freitas L, Antonino JD, Rancurel C, Mota Do Carmo Costa M, Grossi-de-Sa MF, Miller RNG, Brasileiro ACM, Messenberg Guimaraes P, et al. 2020a. Comparative Genomics Reveals Novel Target Genes towards Specific Control of Plant-Parasitic Nematodes. *Genes* 11:1347.

Holterman M, Karegar A, Mooijman P, Van Megen H, Van Den Elsen S, Vervoort MTW, Quist CW, Karssen G, Decraemer W, Opperman CH, et al. 2017. Disparate gain and loss of parasitic abilities among nematode lineages.Wong WO, editor. *PLOS ONE* 12:e0185445.

Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. 2017. WormBase ParaSite − a comprehensive resource for helminth genomics. *Mol. Biochem. Parasitol.* 215:2–10.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* 33:1635–1638.

Ishikawa SA, Zhukova A, Iwasaki W, Gascuel O. 2019. A Fast Likelihood Method to

Reconstruct and Visualize Ancestral Scenarios.Pupko T, editor. *Mol. Biol. Evol.* 36:2069–2085.

Jacob F. 1977. Evolution and Tinkering. *Science* 196:1161–1166.

Jowkar G, Pečerska J, Maiolo M, Gil M, Anisimova M. 2023. ARPIP: Ancestral Sequence Reconstruction with Insertions and Deletions under the Poisson Indel Process.Paterson A, editor. *Syst. Biol.* 72:307–318.

Katoh K. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066.

Kelly S, Maini PK. 2013. DendroBLAST: Approximate Phylogenetic Trees in the Absence of Multiple Sequence Alignments.Robinson-Rechavi M, editor. *PLoS ONE* 8:e58537.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet.* 25:404–413.

Koutsovoulos GD, Poullet M, Elashry A, Kozlowski DKL, Sallet E, Da Rocha M, Perfus-Barbeoch L, Martin-Jimenez C, Frey JE, Ahrens CH, et al. 2020. RETRACTED ARTICLE: Genome assembly and annotation of Meloidogyne enterolobii, an emerging parthenogenetic root-knot nematode. *Sci. Data* 7:324.

Lebherz MK, Fouks B, Schmidt J, Bornberg-Bauer E, Grandchamp A. 2024. DNA Transposons Favor De Novo Transcript Emergence Through Enrichment of Transcription Factor Binding Motifs.Luria V, editor. *Genome Biol. Evol.* 16:evae134.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci.* 103:9935–9939.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323.

Li C-Y, Zhang Yong, Wang Z, Zhang Yan, Cao C, Zhang P-W, Lu S-J, Li X-M, Yu Q, Zheng X, et al. 2010. A Human-Specific De Novo Protein-Coding Gene Associated with Human Brain Functions.Bourne PE, editor. *PLoS Comput. Biol.* 6:e1000734.

Lilley CJ, De Souza VHM, Eves-van Den Akker S. 2024. Plant diseases caused by nematodes. In: Agrios' Plant Pathology. Elsevier. p. 607–649. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780128224298000182

Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G. 2003. PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 17:2337–2342.

McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* 17:567–578.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.Teeling E, editor. *Mol. Biol. Evol.* 37:1530–1534.

Mitchum MG, Hussey RS, Baum TJ, Wang X, Elling AA, Wubben M, Davis EL. 2013a.

Nematode effector proteins: an emerging paradigm of parasitism. *New Phytol.* 199:879–894.

Mota APZ, Koutsovoulos GD, Perfus-Barbeoch L, Despot-Slade E, Labadie K, Aury J-M, Robbe-Sermesant K, Bailly-Bechet M, Belser C, Péré A, et al. 2024. Unzipped genome assemblies of polyploid root-knot nematodes reveal unusual and clade-specific telomeric repeats. *Nat. Commun.* 15:773.

Nevers Y, Rossier V, Train CM, Altenhoff A, Dessimoz C, Glover N. 2022. Multifaceted quality assessment of gene repertoire annotation with OMArk. Bioinformatics Available from: http://biorxiv.org/lookup/doi/10.1101/2022.11.25.517970

Prabh N, Rödelsperger C. 2019. *De Novo* , Divergence, and Mixed Origin Contribute to the Emergence of Orphan Genes in *Pristionchus* Nematodes. *G3 GenesGenomesGenetics* 9:2277–2286.

Qing X, Zhang YM, Sun S, Ahmed M, Lo W-S, Bert W, Holovachov O, Li H. 2024. Phylogenomic Insights into the Evolution and Origin of Nematoda.Struck T, editor. *Syst. Biol.*:syae073.

Schmitz JF, Bornberg-Bauer E. 2017. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. *F1000Research* 6:57.

Seçkin E, Colinet D, Sarti E, Danchin EGJ. 2025. Orphan and *de novo* Genes in Fungi and Animals: Identification, Origins and Functions.Luria V, editor. *Genome Biol. Evol.* 17:evaf220.

Silveira AB, Trontin C, Cortijo S, Barau J, Del Bem LEV, Loudet O, Colot V, Vincentz M. 2013. Extensive Natural Epigenetic Variation at a De Novo Originated Gene.Purugganan MD, editor. *PLoS Genet.* 9:e1003437.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35:1026–1028.

Szitenberg A, Salazar-Jaramillo L, Blok VC, Laetsch DR, Joseph S, Williamson VM, Blaxter ML, Lunt DH. 2017. Comparative Genomics of Apomictic Root-Knot Nematodes: Hybridization, Ploidy, and Dynamic Genome Change. *Genome Biol. Evol.* 9:2844–2861.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12:692–702.

Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, Von Heijne G, Nielsen H. 2022. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* 40:1023–1025.

Thumuluri V, Almagro Armenteros JJ, Johansen AR, Nielsen H, Winther O. 2022. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Res.* 50:W228–W234.

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Mar Alba M. 2008. Origin of Primate Orphan Genes: A Comparative Genomics Approach. *Mol. Biol. Evol.* 26:603–612.

Vakirlis N, Carvunis A-R, McLysaght A. 2020. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* 9:e53500.

Van Megen H, Van Den Elsen S, Holterman M, Karssen G, Mooyman P, Bongers T, Holovachov O, Bakker J, Helder J. 2009. A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. *Nematology* 11:927–950.

Van Oss SB, Carvunis A-R. 2019. De novo gene birth. *PLOS Genet.* 15:e1008160.

Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, Lee T -h., Jin H, Marler B, Guo H, et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49–e49.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10:57–63.

Warwicker J, Charonis S, Curtis RA. 2014. Lysine and Arginine Content of Proteins: Computational Analysis Suggests a New Tool for Solubility Design. *Mol. Pharm.* 11:294–303.

Weisman CM. 2022. The Origins and Functions of De Novo Genes: Against All Odds? *J. Mol. Evol.* 90:244–257.

Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. 2013. Mechanisms and Dynamics of Orphan Gene Emergence in Insect Genomes. *Genome Biol. Evol.* 5:439–455.

Zdobnov EM, Apweiler R. 2001. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847–848.

Zhang W, Gao Y, Long M, Shen B. 2019. Origination and evolution of orphan genes and de novo genes in the genome of Caenorhabditis elegans. *Sci. China Life Sci.* 62:579–593.

Zhang Z. 2022. KaKs_Calculator 3.0: Calculating Selective Pressure on Coding and Non-Coding Sequences. *Genomics Proteomics Bioinformatics* 20:536–540.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in *Drosophila*. *Genome Res.* 18:1446–1455.

Zhu Z, Bo D, Xie C, Dai D, Peng D, Sun M, Zheng J. 2025. Integrative multi-omics analysis reveals the translational landscape of the plant-parasitic nematode Meloidogyne incognita. *Commun. Biol.* 8:140.

Zhuang X, Yang C, Murphy KR, Cheng C-HC. 2019. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *Proc. Natl. Acad. Sci.* 116:4400–4405.