# T2T genome of the basal nematode *Xiphinema index* reveals major adaptive events associated to plant parasitism and epigenetic regulation

Julia Truch[1*], Karine Robbe-Sermesant[1*], Arthur Péré[1*], Dominique Colinet[1], Corinne Rancurel[1], Elodie Drula[2], Martine Da Rocha[1], Laetitia Perfus-Barbeoch[1], Ercan Seçkin[1], Ulysse Julien-Portier[1], Céline Lopez-Roques[3], Carole Iampietro[3], Amalia Sayeh[3], Marie Gislard[3], Roxane Boyer[3], Léo Luvisutto[3], Daniel Esmenjaud[1], Etienne GJ Danchin[1*c] & Cyril Van Ghelder[1*c]

* equal contributions
[c] corresponding authors

## Affiliations

[1]Institut Sophia Agrobiotech, INRAE, Université Côte d'Azur, CNRS, 400 routes des Chappes, 06903, Sophia-Antipolis, France.
[2]NRAE, Aix Marseille Univ, BBF, Biodiversité et Biotechnologie Fongiques, AFMB, Architecture et Fonction des Macromolécules Biologiques, USC1408, Marseille, France.
[3]INRAE, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France.

## Authors contributions

JT: Epigenetic experiments and analyses, contributed to the final manuscript
KR-S: Genome assembly, data analyses,contributed to the final manuscript
AP: Genome assembly, data analyses
DC: HGT predictions and analyses, contributed to the final manuscript
CR: Transcriptome assembly and gene prediction, mitochondrial genome assembly, contributed to the final manuscript, data submission
ED: Predictions and analysis of CAZymes
MDR: Transcriptomic and TE analyses
LP-B: Production of genomic material
ES: Data analyses
UJP: Nematode production
CL-R: Supervised sequencing and contributed to the final manuscript
CI: Performed Hi-C library and sequencing
AS: Bioinformatic pre-treatment of ONT data
MG: Performed PACBIO library and sequencing
RB: Performed ONT libraries and sequencing
LL: Quality controls and sequencing of ATAC libraries
DE: Nematode production and expertise
EGJD: Project coordination, funding acquisition, manuscript writing, data analyses
CVG: Nematode production, Project coordination, production of genomic and transcriptomic material, molecular biology validations, manuscript writing, data analyses, data submission

# Abstract

Nematodes constitute the most speciose animal phylum, occupying an exceptionally broad range of biomes and encompassing free-living microbe-feeders, predators, and notorious plant or animal parasites. However, the limited availability of genome-scale resources for early-branching nematode clades has constrained our understanding of the molecular determinants underlying these multiple evolutionary adaptations. Here, we present the genome of the plant-parasitic nematode *Xiphinema index*, constituting the first telomere-to-telomere assembly for the order Dorylaimida, an early-branching nematode lineage. Phylogenetic analyses reveal that horizontal gene transfers (HGT) have played a pivotal role in the emergence of plant parasitism in this phylum. Remarkably, we identify the first documented case of a plant-derived HGT in a nematode genome. This gene encodes a predicted secreted parasitic effector, underscoring the diversity of evolutionary events that contributed to the parasitic arsenal. Our results suggest that, beyond bacteria and fungi, plants themselves have provided crucial genes for parasitism that may now be used by the nematodes to manipulate them. Our comparative genomic analyses across more than 70 nematode species further uncover an unprecedented epigenetic toolkit in an invertebrate genome. Indeed, *X. index* harbors a complete set of canonical DNA methyltransferases, a full-length ATRX chromatin remodeler, and 3D-chromatin architectural CTCF protein coding homolog. Consistent with this repertoire, we detected methylation at CpG genome-wide, which inversely correlates with gene expression and chromatin accessibility. Together, these findings reveal a vertebrate-like epigenetic machinery in an early-branching nematode and call for a fundamental reassessment of the prevailing epigenetic paradigm in nematodes and, more broadly, in animals.



**Keywords:** Plant Parasitic Nematode (PPN), Virus Vector (GFLV), T2T Genome, Horizontal Gene Transfer (HGT), DNA Methyl Transferase (DNMT), ATRX, DNA methylation.

# Introduction

Nematodes are particularly species-rich[1] and are considered the most abundant animals on Earth[2,3]. Despite their marine origin[4–6], they are also present in terrestrial and freshwater environments as a sign of their adaptive evolutionary success. Their ecology is diverse and encompasses free-living microbe-feeders, predators, and notorious plant or animal parasites. This diversity of trophic habits results from multiple independent evolutionary transitions. The limited availability of chromosome-scale genome resources for early-branching nematode clades has constrained our understanding of the molecular determinants underlying these multiple adaptations.

This is the case for plant parasitism that has emerged independently at least four times during the evolution of nematodes[5,7]. Convergent morphological evolution has accompanied the multiple transitions from a free-living to a plant-parasitic lifestyle[8]. This includes the systematic presence of a needle-like stylet in the mouth part used to secrete effector proteins and withdraw nutrients from the plant. The morphology and ontogeny of the stylets are different in each clade of plant-parasitic nematodes, suggesting this organ has evolved independently and convergently. At the molecular level, acquisition of plant cell wall-degrading enzymes via horizontal gene transfers has been considered a key evolutionary step towards adaptation to plant parasitism. Indeed, cellulases and other enzymes have been acquired, mostly from bacteria, in major plant parasites such as cyst and root-knot nematodes, while similar enzymes in foliar nematodes have fungal origins[9–11]. However, most of our knowledge on the molecular determinants underlying adaptation to plant parasitism is restricted to these two relatively recent nematode clades as no substantial genome resource is available in basal clades of plant parasites. Consequently, whether evolutionary events such as HGT or others are systematically associated with adaptation to plant parasitism in nematodes, including in basal clades, is yet to be determined. Preliminary transcriptomic data in *Xiphinema index* and *Longidorus elongatus* suggest HGT events took place in the basal clade of plant-parasitic nematodes to which they belong as well[12]. However, in the absence of a genome, distinguishing HGT events from contamination in transcriptomic data alone is barely possible. Despite efforts recently deployed to produce genomes in basal nematode clades[13], no chromosome-level genome assembly is available for a plant parasite in those clades so far.

Therefore, to better understand the evolutionary events associated with adaptation to plant parasitism and shed light on possible other singularities, we have produced a high-quality T2T chromosome-scale genome for the dagger nematode *Xiphinema index*, as a representative of plant parasites in the basal order Dorylaimida. *X. index* Thorne & Allen, 1950 is a virus vector and migratory ectoparasite dreaded by the viticulture sector. This species feeds on actively growing grapevine roots causing malformation, swelling, galling made of multinucleate cells, or necrosis[14,15]. Although high inoculum pressure of *X. index* may cause growth retardation to host plants, the severity of its impact on grapevine, *Vitis vinifera*, is mostly due to its ability to transmit the Grapevine Fanleaf Virus (GFLV)[15]. This nepovirus is specifically transmitted by *X. index* and is the main cause of the lethal grapevine degeneration disease, making it the most severe viral disease of grapevines worldwide[16]. In France, one third of the vineyards are severely affected by the grapevine degeneration disease[17]. Native from the middle east, *X. index* was dispersed later along with grapevine's agricultural history, and is nowadays present in most if not all vineyards worldwide[18]. Consequently, *X. index* was the only virus vector to be included in the top 10 plant parasitic

nematodes (PPN) in terms of agronomical and scientific importance[19]. At an agricultural point of view, while both GFLV variants and *Vitis* species and cultivars have now a significant number of omics resources available, these crucial assets were lacking for *X. index* impeding the in-depth study of this three-partners pathosystem.

Comparative analysis of the high-quality *X. index* genome with those of 70 other nematode species, combined with large-scale phylogenetic analyses revealed several major findings. First, we confirmed plant cell wall-degrading enzymes and other proteins were also acquired by HGT in *X. index*, suggesting these evolutionary events played major roles in the adaptation to plant parasitism across the nematode tree of life. Remarkably, in addition to HGT of bacterial and fungal origin, we also identified the first case of HGT of plant origin in a nematode. The identified gene of plant origin encodes a putative secreted calmodulin with potential role in plant-nematode interaction.

Second, our comparative analysis revealed in *X. index*, an epigenetic toolkit of unparalleled completeness in nematodes; and more broadly in invertebrates. In most eukaryotes, Cytosine methylation (5mC) of CG dinucleotides is a fundamental modification of the epigenetic landscape[20,21]. In vertebrates, it is commonly accepted that 80% of CG dinucleotides are methylated. Among others, CpG methylation at the promoter regions is often linked with gene repression. Deposition and maintenance of 5mC is primarily orchestrated by the interplay of DNA methyl transferase (DNMTs) enzymes. Generally, DNMT3 is responsible for *de novo* DNA deposition of CpG, whereas DNMT1 is involved in the maintenance of these epigenetic marks. DNMT2, however, is often associated with RNA methylation[22]. By contrast, invertebrates, which represent the vast majority of animals, have a differing DNA methylation pattern and machinery. For instance, nematodes were so far considered to have a null or sparse methylation on Cytosine (5mC) across genomes and often lack some components of the machinery required for the deposition and/or maintenance of the later one[23–27]. Indeed, the presence/absence of canonical DNMTs is sparse among nematode genomes. Other key factors of the epigenetic machinery are essential for the maintenance of the epigenetic landscape, among which the chromatin remodelers such as ATRX (alpha-thalassemia mental retardation X-linked) [28]. This member of the SWI/SNF family acts as an ATP-dependent molecular motor and interacts among others with DNA tandem repeat sequences in both telomeres and euchromatin, regulatory elements and DNA G-quadruplex structures (G4)[29]. ATRX mutations or knock-down affects important processes including DNA repair and methylation and gene expression. In vertebrates, ATRX is composed of two main domains, a N-terminal ADD domain and a C-terminal helicase/ATPase domain[30]. Its orthologs, named Xnp, in *Drosophila melanogaster* and *C. elegans* are truncated and deprived of ADD domain. Interestingly, the characteristics of this domain have been only found in DNMT3 proteins all involved in DNA methylation and the latter two species show rare or absent DNA methylation[31]. In the *X. index* genome, we identified (i) a full and large arsenal of DNMT, (ii) a global genome methylation at CGs and (iii) an anti-correlation between 5meC levels at TSS and gene expression and accessibility of the chromatin. Furthermore, we identified the first case of an ATRX protein with full-length ADD domain in invertebrates. These unexpected findings suggest similarities between the epigenetic arsenal of this nematode and vertebrates[32] and impose a rethinking of the evolution of the epigenetic toolkit across the animal tree of life.

# Results

## The first T2T reference genome for nematodes of the Dorylaimida order

### Ten chromosomes hosting abundant coding genes

Using a combination of Nanopore, PacBio HiFi and Hi-C reads, we obtained three initial assemblies, one for each of the randomly separated haplotypes (hap1 and hap2), and one merged assembly representing a consensus of both haplotypes (merge). Comparative analysis of the k-mers in the HiFi reads and in the assemblies showed that the sequence information present in the reads was well captured by our assemblies. The two haplotypes seem to be divergent enough to contain a substantial portion of haplotype-specific k-mers as illustrated by the dark peak of k-mers found 0x in the assembly but at the same multiplicity than the 1X heterozygous peak on the k-mer plots **(Figure 1)**. Concatenating the two haplotypes in a pseudo-diploid assembly leads to the elimination of the dark peak thus confirming this hypothesis. The estimated heterozygosity of ~0.9% seems sufficient to explain the observed patterns in k-mer plots **(Figure 1)**. After scaffolding with Hi-C data, gap-filling and elimination of contamination, the merged assembly counted 73 contigs with a N50 length of 23.66 Mb and a L50 of 5, for a total size of 225.83 Mb (**Table S1**). This final *Xiphinema index* genome assembly (merged) will be used as a reference for the following analyses. The genome size, larger than those observed in other Dorylaimia nematodes (50-100 Mb in Trichinellida and 140Mb in Mesodorylaimus), is consistent with the genome size estimated using k-mer (ca. 219 Mb haploid) **(Figure 1).** We found that the canonical nematode telomeric sequence (TTAGGC)n formed repeat arrays at both ends of the 10 biggest scaffolds suggesting a telomere-to-telomere (T2T) assembly of 10 chromosomes, which is consistent with previous cytogenetic observations[33].

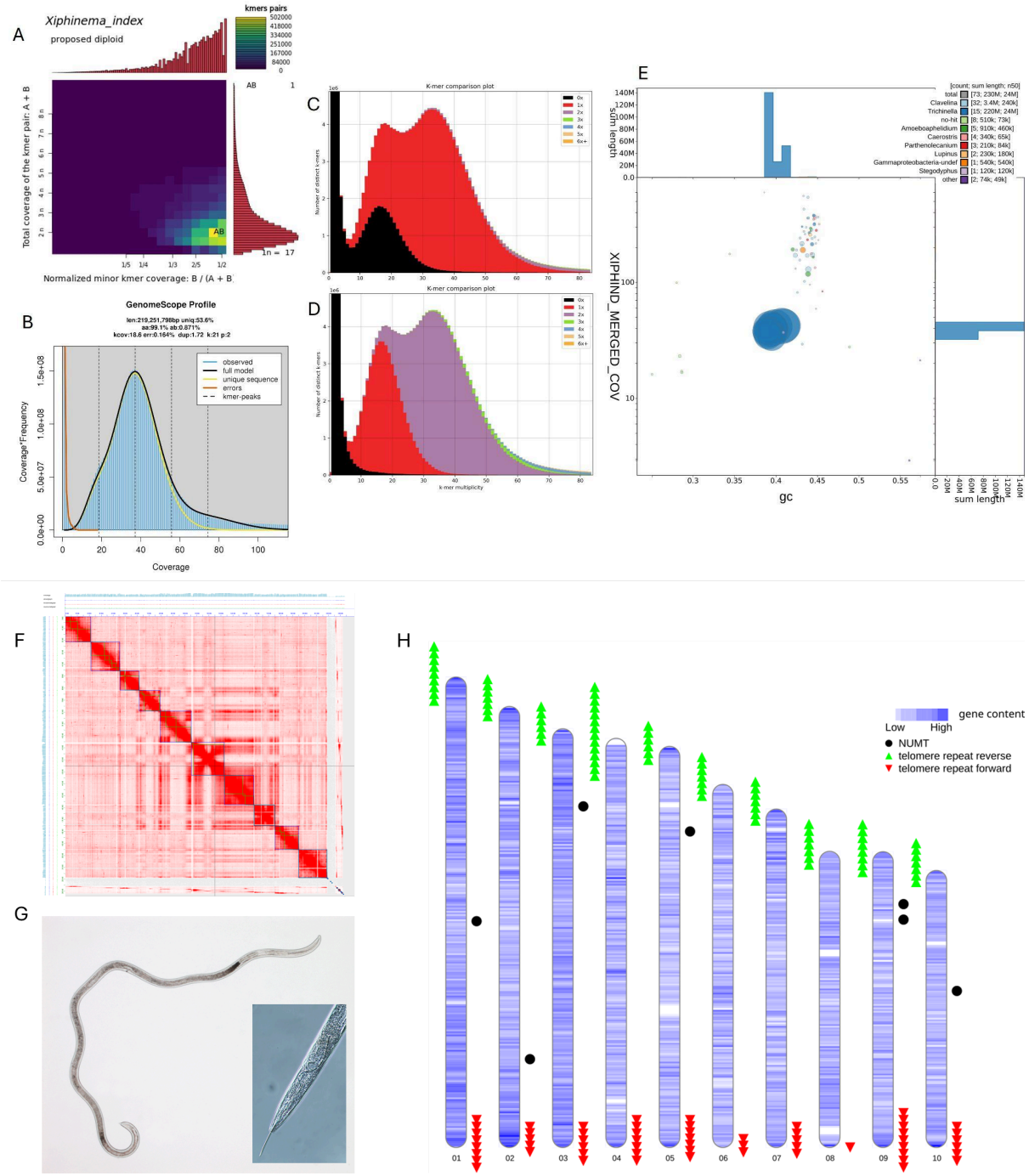



**Figure 1:** Genome profiling of *Xiphinema index*. (A) Smudgeplot of *X. index* extracting 21-mers from HiFi reads. The color intensity of each smudge reflects the approximate number of k-mers per bin. *X. index* is predicted to have a diploid genome. (B) GenomeScope2 k-mer profile and estimated parameters for *X. index*. Coverage (kcov), error rate (err.), haploid genome size estimation (len.), k-mer size (k) and ploidy level (p). *X. index* shows an average heterozygosity of ca. 0.9% with a predicted haploid genome size of 219 Mb. (C) Genome assembly k-mer spectra. The Kat spectrum plot using HiFi reads reports the multiplicity of each k-mer detected in the read set and in the assembly. The plot is color-coded according to the number of times a k-mer is found in the assembly. The plot represents the merged consensus assembly. (D) The plot represents concatenation of haplotypes 1 and 2. (E) BlobPlot of *X. index* genome assembly (merged) showing taxonomic affiliation at the genus rank level. Sequence length (sum) is mentioned above and on the right panel. Marginal contamination is detected in debris. (F) Genome-wide Hi-C contact matrix of *X. index*. (G) *X. index* adult female 3.5mm long with a close-up of its large stylet (classic apparatus in PPN). (H)

representation of the 10 chromosomes of *X. index* showing the gene density, mitochondrial DNA insertions and positions of oriented telomeric repeats.

A total of 33,647 genes have been predicted, of which 30,614 are protein-coding genes. The number of protein-coding genes is higher than observed in other Dorylaimia species (13-16,000 for *Trichinellida* and ca. 23,000 in *Mesodorylaimus*, **Table S2**). The functional annotation revealed that 54.6% of the *X. index* predicted proteins contained at least one InterPro domain, covering 10,845 different domains in total. The InterPro domains and Gene Ontology annotations present in *Xiphinema index* and absent in other studied members of the Dorylamia family are listed in **Tables S3** and **S4,** respectively. The full lists of InterPro domains and GO annotations that show a significantly higher representation in *Xiphinema index* over the other Dorylamia are summarized in (recherche data gouv). A total of 2959 proteins have a predicted signal peptide for secretion but neither predicted transmembrane region nor GPI anchor. These proteins represent the predicted secretome of *X. index* and will be referred to as PSP for predicted secreted proteins in the rest of the manuscript.
To assess the completeness of *X. index* genome, we calculated and compared the BUSCO metazoa completeness at the genome and predicted proteome levels, to five other nematode species belonging to the Dorylaimia clade. With 84.5% and 86.9% of complete BUSCO genes retrieved, at the genome and predicted proteome levels, respectively, both the genome and annotation of *X. index* show a substantially higher completeness than those of the animal-parasitic *Trichinellidae* genomes and a completeness comparable to that of *Mesodorylaimus*, which is a free-living species (**Table S2**).

## Mitochondrial insertions into the nuclear genome

The mitochondrial genome of *X. index* is 14,820 bp long with a GC content of 30.41%. This mitogenome is substantially larger than those of the *Xiphinema americanum* group *sensu lato*, *X. americanum* (NC_005928.1 - 12,626 bp), *X. rivesi* (NC_033869.1 - 12,624 bp) or *X. pachtaïcum* (NC_033870.1 - 12,489 bp) **(Figure S1)**. It contains 40 genes including 12 protein coding genes (all the canonical genes except *ATP8*), 2 rRNAs, 26 tRNAs and 2 non-genic regions that are reverse complementary to each other. Some conserved suite of genes between *X. index* and *Longidorus vineacola* (NC_033867) or *Paralongidorus litoralis* (NC_033868) were identified (**Figure S2**) [34]. The two duplicated non-coding regions (NCR) mostly explain the larger size of the mitogenome of *X. index.* This curiosity is confirmed by alignment of Oxford Nanopore Technologies (ONT) reads onto the mitogenome. Using these regions as BLAST queries, no homology with any other sequences in public libraries was found. We analyzed 25 full-length mitochondrial genomes available in Dorylaimia and confirmed that *ATP8* is only found in Trichinellida. The mitogenome size of *X. index* is within the Dorylaimia range (min-max: 12,489bp in *X. pachtaicum* - 26,194bp in *Romanomermis culicivorax*) and the observed size variations are directly correlated ($R^2$= 0.94) with the size of the non-genic regions, which are present in more than half of the mitogenomes (**Figure S1**).
Alignments between nuclear and mitochondrial genomes coupled with visualization revealed several insertions of mitochondrial genetic material into the nuclear genome or nuclear mitochondrial DNA segments (NUMTs). Insertion positions were determined by BLAST search. A total of 12 NUMTs (>500 bp) were found within the chromosomes **(Figure 2**). These results were supported by overlapping long reads. Most of these insertions are heterozygous (present in one or the other haplotype) including the largest one which is 11.4

kb long on the chromosome 10 (K10 insertion). This latter insertion, which is almost as long as the whole mitochondrial genome and 96.24% identical, was confirmed by PCR. PCR primers that flank the junctions between the nuclear genome and the mitochondrial transferred region, successfully amplified the expected junction site, at both ends of the insertion. PCR products, which were Sanger sequenced, perfectly aligned onto the genome region, thus validating the assembly. These primers were used on an unrelated population of *X. index* originally from Iran (*X. index "Iran"*), available in the lab nematode collection. Interestingly, it gave the same result with a slight difference at the insertion point (i.e. tiny indel at both ends, **Figure 2**). The mitochondrial genes inserted in the nuclear genome were not annotated by the nuclear gene predictor and no substantial transcriptional support was detected in this region, suggesting that they are inactive. In Dorylaimia, three other chromosome scale genome assemblies with the corresponding mitochondrial genome are available: *Trichinella spiralis, T. murrelli and Echinomermella matsi.* BLAST analysis carried out on these datasets failed to detect any NUMT in these species (**Figure S1**).

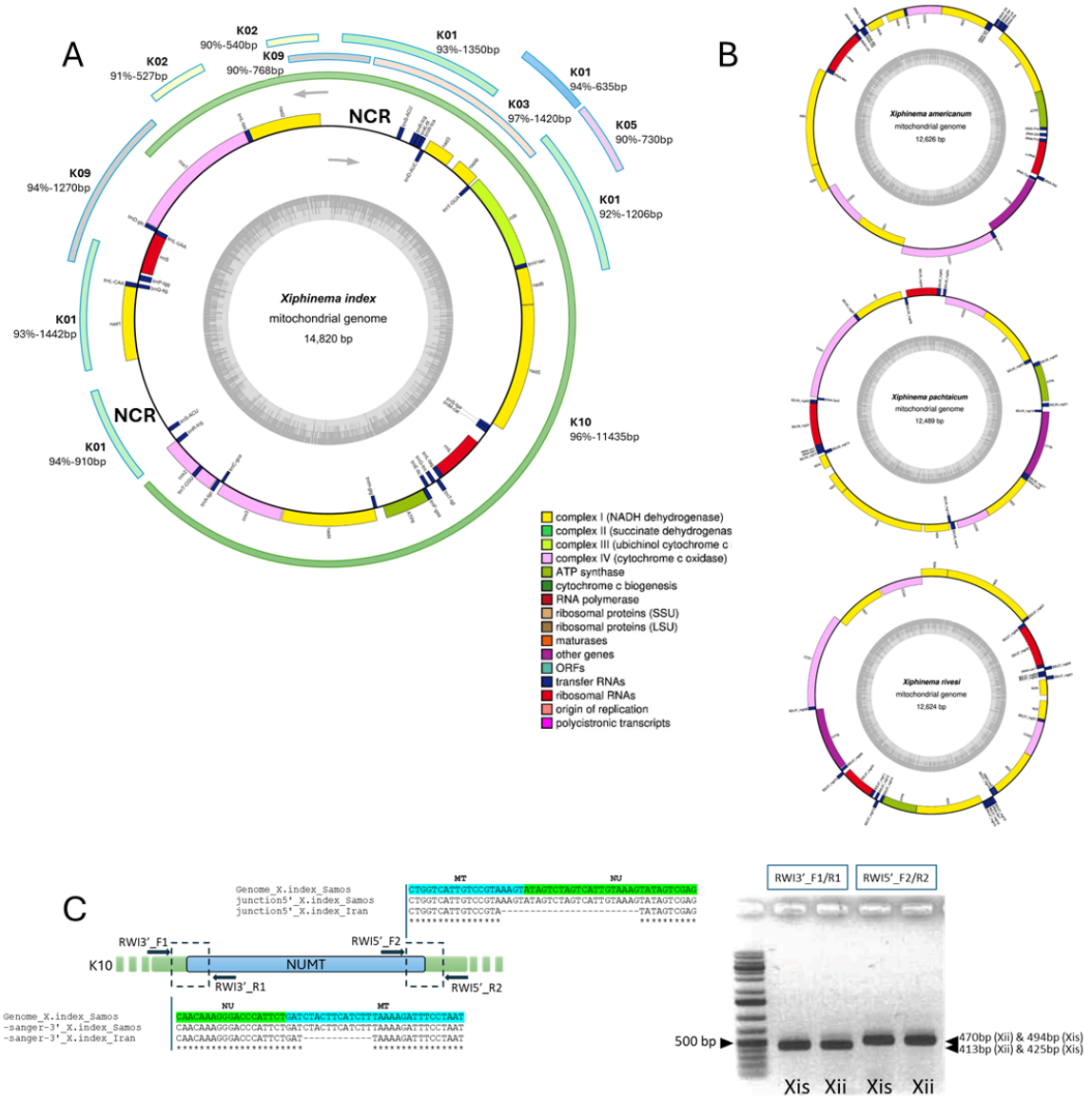


**Figure 2:** Mitochondrial genome and NUMTs of *Xiphinema index.* A) mitochondrial genome of *X. index* with the NUMTs positioned on the mitogenome. Their chromosome locations (K1 to K10), lengths and percentages of identity with the corresponding sequences in the mitogenome are indicated. NCR for non coding region. B) Mitogenomes of *X. americanum* (NC_005928.1), *X. rivesi* (NC_033869.1), and *X. pachtaïcum* (NC_033870.1). C) Validation of the main NUMT on the chromosome 10 (K10). Two distinct populations of *X. index* were PCR-tested with 5' and 3' flanking primers (RW5' and RW3') and sanger sequenced: *X. index* 'samos' (Xis) corresponding to the genome (Xis) and *X. index* 'Iran' (Xii).

## *X. index* chromosomes have canonical telomeric repeat but encode only a partial repertoire of telomere-associated proteins

Variations around the canonical nematode telomere system have been identified across different nematode clades, including in Dorylaimia with Trichinellida as sole representatives of this clade by that time [35]. Therefore, we took advantage of the available annotated genome of *Mesodorylaimus sp. YZB2_4* together with the one of *X. index* to further explore the telomere systems in the Dorylaimia clade. We complemented the previous Nematoda telomeres comparative genomics analysis [35] with these two Dorylaimida genomes and used the same criteria to determine the presence / absence patterns of telomerase and other telomere-associated proteins in nematode genomes. In addition to the presence of a canonical TTAGGC nematode telomeric DNA repeat in *X. index* and *Mesodorylaimus*, we identified a candidate telomerase (xiphind_chr_06g013610 for *X. index*) enzyme in both species, confirming a general trend of co-occurence of the repeat and the enzyme. However, the proteins composing the shelterin complex, which is associated with telomeres, showed more variations in terms of conservation. Only partial evidence for orthologs of the *C. elegans* double strand-telomeric DNA binding proteins tebp1 and tebp2 was found in *X. index* and no evidence at all in the rest of the Dorylaimia. Similarly, the single-strand-telomeric DNA binding proteins pot1, pot2 and pot3 had no identifiable homolog in any of the Dorylaimia species analyzed. However mrt1, another single-strand-telomeric DNA binding protein was conserved in both *X. index* and *Mesodorylaimus* as well as the other Dorylaimia, except two Trichinella species. Curiously, Mrt2, which has a dual role with Clk2 in telomere length regulation and DNA damage checkpoints in *C. elegans* is missing in *X. index* while it is well conserved in the rest of Dorylaimia nematodes including *Mesodorylaimus*. **(Figure 3)**.

Overall, this comparative genomics analysis suggests that although a simple TTAGGC telomeric repeat and a telomerase enzyme were probably ancestrally present in nematodes, multiple independent losses of this system took place, including in the Trichinellidae among the Dorylaimia. Conservation of proteins that bind telomeric DNA between the model nematode *C. elegans* and the other nematodes is even more disparate, suggesting either rapid evolution of these proteins or recruitment of different proteins to play the same role in different nematodes.

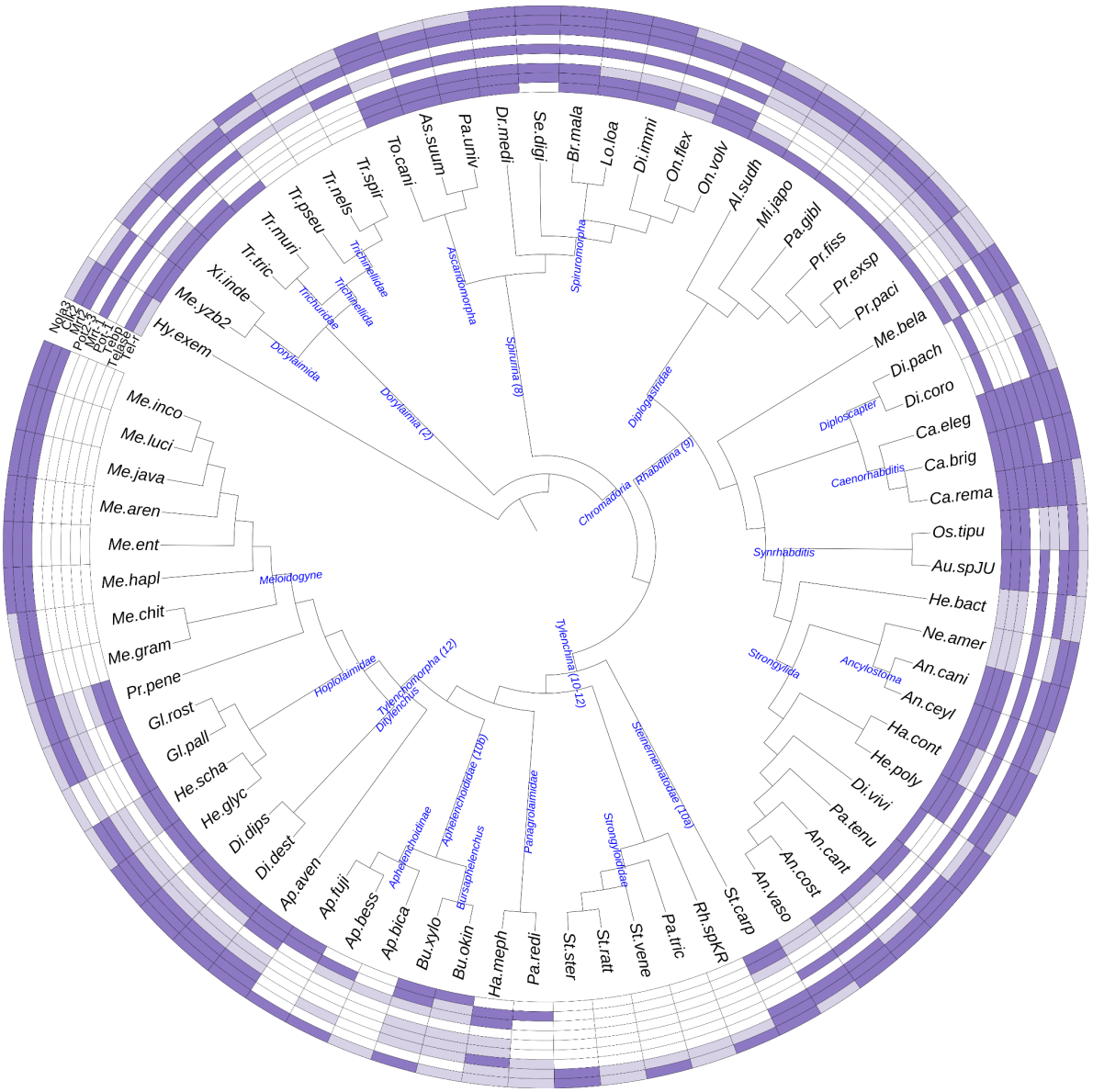



**Figure 3:** Distribution of telomere repeats and telomere-associated proteins in the phylum Nematoda. Violet color means present, light-violet means partial evidence and blank no evidence. Tel-r: *C. elegans* telomeric repeat (TTAGGC)n except for the outgroup *Hypsibus exemplaris* in which another simple 9-nucleotides repeat has been described [36]. Telase: telomerase reverse transcriptase. Tebp: ds-telomeric DNA binding proteins Tebp1 or Tebp2. Pot1: ss-telomeric DNA binding protein Pot-1. Mrt1: ss-telomeric DNA binding protein Mrt-1. Pot2-3: ss-telomeric DNA binding proteins Pot2 or Pot3. Mrt2 and Clk2: proteins with a putative function in telomere length regulation. Nola3: H/ACA ribonucleoprotein complex subunit 3.

# Horizontal gene transfers (HGT) contribute to the repertoire of *Xiphinema*-specific proteins, CAZymes and predicted secreted proteins (PSP)

## Xiphinema-specific proteins

To identify the degree of conservation vs. genus-specificity of *X. index* predicted proteins, we compared them against those of 71 other nematodes and one tardigrade outgroup species using Orthofinder. Of the 30,614 predicted proteins in *X. index*, 11,564 (38%) had no predicted ortholog in the other nematode proteomes used in the analysis and thus constitute candidate Xiphinema-specific (X-s) proteins. It was the highest rate of proteins unassigned to an orthogroup among all the other nematodes included, probably because *X. index* is the sole representative of the Longidoridae family in this analysis. The highest number of shared orthologues was 16,531, with *Mesodorylaimus sp. YZB2_4*, the closest Dorylaimida species included in our analysis. In comparison, the number of orthologues conserved with Trichinella spp. or Trichuris spp. was between ca. 8000 and 9000, suggesting extensive gene losses in the Trichinellidae. Among the X-s predicted proteins, less than 10% (1024 predicted proteins) had at least one predicted Interpro domain **(Figure S3)**. From these 1024 X-s predicted proteins, 16% were related to ubiquitins among which the large majority (149 genes) are associated with predicted TE, 13% to peptidase of diverse types, and 11% of Zinc finger-containing proteins **(Figure S3)**. A total of 40 predicted proteins related to G protein-coupled receptors had no predicted ortholog in the other nematode proteomes tested. These proteins contained GPCR rhodopsin-like or secretin, arrestin or serpentine domains and represent 4% of the corresponding protein set identified in the whole proteome of *X. index* **(Figure S3)**. As this large gene family is usually involved in the sensory machinery, X-s GPCR-related proteins might be important for this ectoparasite moving in the soil to sense its environment including potential host plants.

## A CAZome composition reflecting the ability to degrade plant carbohydrates

In the context of plant-parasite interactions, Carbohydrate-Active enZymes (CAZymes) play an essential role in degrading the plant cell wall and digesting storage sugars. It is therefore crucial to delineate their repertoire. A total of 392 Carbohydrate-Active enZymes (CAZymes) were identified in the set of predicted *X. index* proteins (**Table S2**). These CAZymes contained 473 modules, including 115 Glycoside Hydrolases (GH), 284 Glycosyl Transferases (GT), and one Carbohydrate Esterase (CE), but no Polysaccharide Lyase (PL) **(Table 1)**. Although PL enzymes are usually absent from animal genomes, some pectate lyases involved in pectin degradation have been identified in other plant-parasitic nematodes, including the pine wilt disease agent *B. xylophilus*, root-knot nematodes and cyst nematodes [10,11]. In addition to CAZyme catalytic modules, carbohydrate-binding modules were identified in 59 *X. index* proteins as well as 14 modules with auxiliary activities associated with CAZymes.

We compared the CAZome (set of identified CAZymes) of *X. index* to those of five other nematode species in the Dorylaimia group, which were selected based on their BUSCO completeness. These included four Trichinellida species that parasitize animals: *Trichinella spiralis* [37], *Trichinella nelsoni* [38], *Trichuris muris* [39] and *Trichuris trichiura* [39], as well as the free

living species *Mesodorylaimus sp. YZB2_4* [40]. The latter is the only species in the Dorylaimida, which is the same group as *X. index*, with a publicly available predicted proteome of good completeness.

| Species / CAZyme modules | GH | GT | PL | CE | CBM | AA | Total modules |
|---|---|---|---|---|---|---|---|
| *Xiphinema index* | 115 | 284 | 0 | 1 | 59 | 14 | 473 |
| *Mesodorylaimus sp. YZB2_4* | 132 | 250 | 0 | 2 | 74 | 14 | 472 |
| *Trichinella nelsoni* | 88 | 139 | 0 | 4 | 45 | 1 | 277 |
| *Trichinella spiralis* | 43 | 88 | 0 | 1 | 22 | 1 | 155 |
| *Trichuris muris* | 57 | 124 | 0 | 3 | 59 | 3 | 246 |
| *Trichuris trichiura* | 61 | 124 | 0 | 2 | 55 | 1 | 243 |

**Table 1**: Annotation of CAZyme modules in the predicted proteomes of nematodes in the Dorylamia group. Modules Abbreviations GH: Glycoside Hydrolase, GT: Glycosyl Transferase, PL: Polysaccharide Lyase, CE: Carbohydrate Esterase, CBM: Carbohydrate Binding Module, AA: Auxiliary Activity. Note that a protein can bear multiple CAZyme modules.

The number of CAZymes was relatively lower in the four Trichinellida species than in the two Dorylaimida species **(Table 1)**. This may reflect a diet less dependent on carbohydrates for the animal parasites as compared to plant parasites or free-living omnivorous species.
Clustering the species according to the abundance in each CAZyme family module recapitulates the phylogeny of the species, with the Trichinellida on one side, and the Dorylaimida on the other side (**Figure S4**).

At the CAZyme family level, we identified three GH enzyme families found only in *X. index*: GH12, GH28 and GH32. The biochemical activities known so far in the GH12 CAZyme family are related to xyloglucan and cellulose degradation. Cellulose is a major component of the plant cell wall and a functional GH12 cellulase has been previously characterized in *X. index* [12]. In the GH28 family, the associated known activity is polygalacturonase, which is involved in pectin degradation, another important component of the plant cell wall. Finally, the associated activity in the GH32 family is fructosidase/invertase. Invertase enzymes degrade sucrose, the major circulating sugar in plants, in glucose and fructose, which can be readily assimilated by animals. All these activities relate to degradation or digestion of plant carbohydrates, therefore we also classified Dorylaimia species according to their abundance of CAZymes in which at least one of the possible activities include degradation of plant sugars (**Figure S5**). In this classification, *X. index* stands out from the rest with a clear outgroup position, reflecting its rich enzymatic arsenal for plant sugar degradation. Corroborating these results, the GO term ‘polygalacturonase activity’ as well as the Interpro domains ‘Glycoside hydrolase, families 28 and 32’ were only found in *X. index* in the comparative InterProScan analysis (**Tables S3** and **S4**).

Some CAZyme families were conserved in the two Dorylaimida species but absent in the Trichinellida species, including GH families 1, 17, 22, 25, 59, 65, 79, 84, and 85, as well as GT families 11 and 109.

Beyond GH12, GH28 and GH32 enzymes with possible activities on plant sugars, we identified a GH31_2 enzyme that is found exclusively in *X. index.* However, no enzymatic activity has been characterized so far in this subfamily. Ultimately, 70 CAZymes are predicted to be secreted, among which 5 are *Xiphinema* specific.

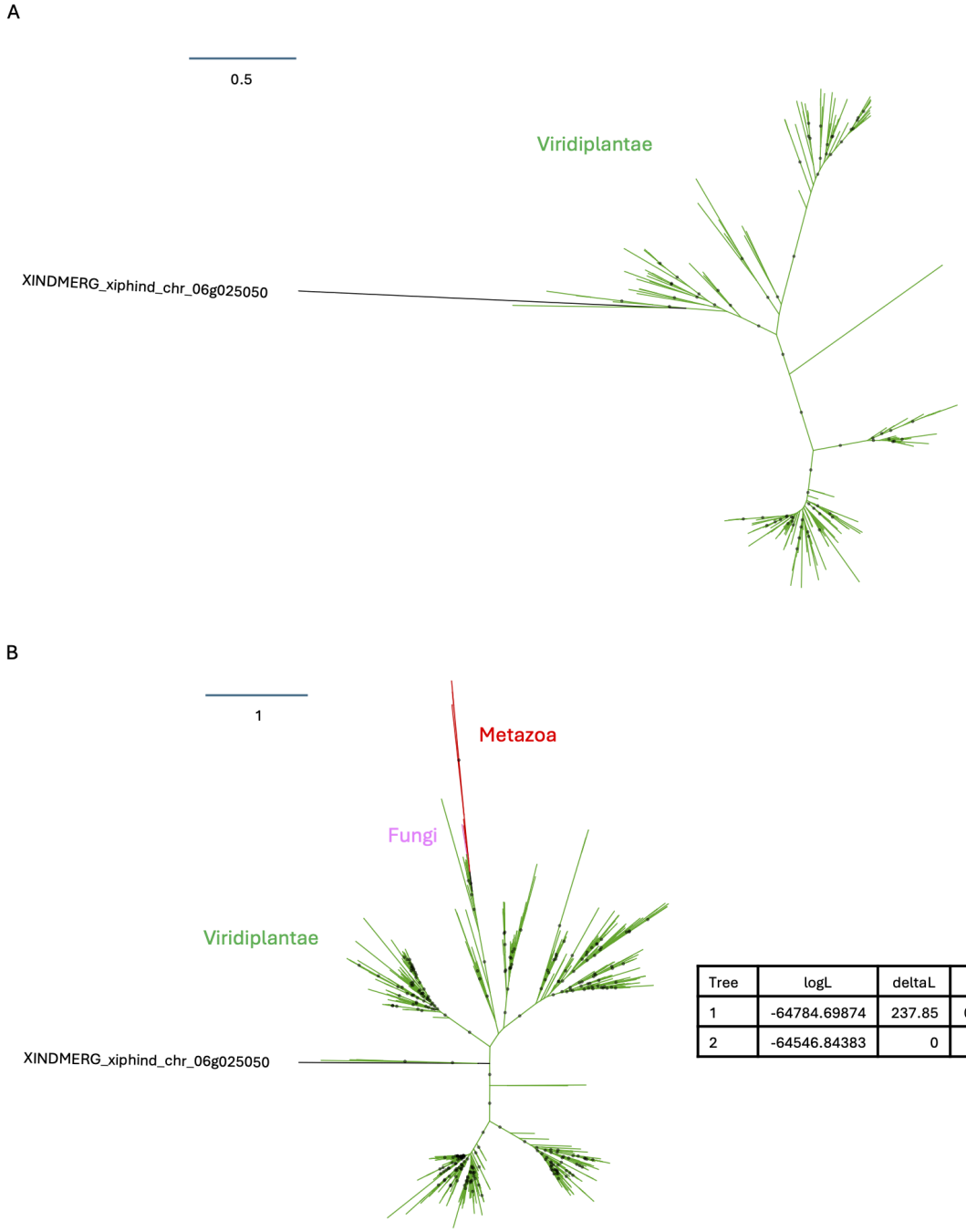
## Acquisition of important enzymes via horizontal gene transfers

Comparative analyses of the predicted proteome with those of other nematode species revealed several *Xiphinema*-specific proteins, including CAZymes, that had no homolog in the rest of nematodes. Furthermore, previous analyses have shown that a GH12 cellulase CAZyme was likely acquired via horizontal gene transfer in *X. index* [12]. To investigate whether HGTs have contributed to the set of X-s proteins and to the CAZome, we performed a comprehensive and accurate phylogenetic detection of HGT candidates, following an approach similar to that described in [41]. Among the 30,614 *X. index* predicted proteins, 1,449 had an Alien Index (AI) or Aggregate Hits Score (AHS) greater than 0, indicating that they are more similar to non-animal than to animal proteins (**Table 2**). Of these, AvP identified 224 as possible Longidoridae-specific HGT candidates. After application of stringent phylogenetic criteria (see Materials and Methods), 44 HGT candidates were validated. These 44 validated HGTs originated from at least 29 independent transfer events, mainly from putative bacterial donors, while a smaller proportion was presumably acquired from fungal donors, and a single case attributed to a plant donor (xiphind_chr_06g025050, **Table 2** and **Table S4**). This atypical HGT is a putative calcium binding protein (calmodulin), while various enzymes represent the large majority of the rest of the HGT. The functional annotation of xiphind_chr_06g025050 revealed a signal peptide for secretion followed by a consensus disordered predicted region and the canonical 4 EF-Hand domains found in calmodulin proteins that are known to bind calcium ions. Using a genomic marker HGT1 designed in this gene, we successfully amplified it in an independent *X. index* population, thus ruling out an hypothetical contamination **(Figure S6)**. As far as we know, this represents the first documented case of HGT of plant origin in a nematode [42]. Accordingly, no Metazoa sequences were found in the phylogenetic tree obtained from a Diamond analysis with the default maximum of 500 hits (**Figure 4A**). Five Metazoa sequences, including two Nematoda sequences, were found in the phylogenetic tree when the Diamond analysis was performed with a maximum of 5,000 hits (**Figure 4B**). However, the five Metazoa sequences were not grouped with the *X. index* sequence. An alternative topology test, constrained to support the monophyly of all Metazoa sequences, was significantly less likely than the unconstrained topology, supporting a plant HGT origin for the *X. index* sequence **(Figure 4B**).
Additionally, we identified six CAZymes likely acquired via HGT in four independent events, all originating from bacteria (**Figure 5** and **Table S5**). These CAZymes belong to three families with known activities on plant sugars: GH12 (cellulase), GH28 (polygalacturonase), and GH32 (invertase) as well as the GH31_2 subfamily of unknown function. Two of the three GH12 are situated 17 kb apart and share 74.6% of sequence identity, thus suggesting a proximal gene duplication event. All of these CAZymes were found to be specific to *X. index* in the previous section on the CAZome analysis. Therefore, HGT events have contributed to the peculiar composition of the *X. index* CAZome. Furthermore, this new comprehensive and stringent analysis confirms the bacterial origin of the GH12 cellulase previously characterized in this species. While candidate GH32 enzymes were also previously reported in *X. index*, no GH28 was identified before in the previous transcriptome analysis of *X. index* and *Longidorus elongatus* [12]. Cross-referencing the Orthofinder

comparative analysis comprising 71 nematode proteomes, with the list of 44 validated HGT cases, returned only 6 cases in common, including the three GH12 sequences **(Figure 6)**. However, a thorough analysis of the 29 phylogenies obtained for the 44 HGT candidates revealed that only four of them are not X-s. These correspond to the XindB02 and XindB11 HGT events, in which sequences from the Dorylaimia *Soboliphyme baturini* are grouped with the *X. index* sequences. The HGT candidates that we confirmed as having originated from Longidoridae-specific HGT of bacterial origin, but that were not in the list of Xiphinema-specific proteins include the sequences belonging to GH28 and GH32 families. These families have also been acquired via HGT in root-knot and cyst nematodes, two distantly related groups of plant-parasitic nematodes [10]. In the absence of the donor species in the OrthoFinder analyses, these proteins are by default reciprocal best hits, showing the limitations of these approaches in such cases of convergent HGT.
Finally, 21 out of 44 HGT are predicted to be secreted making them noteworthy effectors to further study (i.e. PSP are significantly enriched in the HGT subset compared to the whole proteome; Fisher's exact test, $p = 8.5 \times 10^{-11}$) **(Figure 5)**.

| | *Xiphinema index* |
|---|---|
| Number of protein-coding genes | 30,614 |
| Number of predicted proteins with AI or AHS > 0 | 1,449 |
| Number of HGT candidates detected by AvP | 224 |
| - Bacteria | 93 |
| - Fungi | 38 |
| - Viridiplantae | 30 |
| - Other | 63 |
| Number of validated HGT candidates | 44 |
| - Bacteria | 36 |
| - Fungi | 7 |
| - Viridiplantae | 1 |
| Number of validated HGT events | 29 |
| - Bacteria | 23 |
| - Fungi | 5 |
| - Viridiplantae | 1 |

**Table 2.** Overall analysis of HGT candidate search results




| Tree | logL | deltaL | p-AU | |
|---|---|---|---|---|
| 1 | -64784.69874 | 237.85 | 0.00165 | - |
| 2 | -64546.84383 | 0 | 0.998 | + |

**Figure 4 : Phylogenetic analysis of the HGT of plant origin.** (A) Phylogenetic tree from a Diamond analysis with a maximum of 500 hits, as described in the Materials and Methods section. The *X. index* sequence is shown in black. Green branches represent plant sequences. Gray circles indicate nodes with support values greater than or equal to 80% and 90% for SH-aLRT and UFboot, respectively. (B) Phylogenetic tree from a Diamond analysis with a maximum of 5,000 hits and the use of a gap threshold of 0.5 for trimal instead of the -automated1 option. The *X. index* sequence is shown in black. Green branches represent plant sequences. Red branches represent Metazoa sequences. Pink branches represent fungi sequences. Gray circles indicate nodes with support values greater than or equal to 80% and 90% for SH-aLRT and UFboot, respectively. The table corresponds to the result of the approximately unbiased (AU) alternative topology test for the monophyly of metazoan sequences. Tree 1 corresponds to the constrained topology forcing monophyly of metazoan sequences and Tree 2 corresponds to the unconstrained topology. The “+” sign denotes the 95% confidence sets, while the “-” sign denotes significant exclusion. All tests performed 10,000 resamplings using the RELL method.

logL: log-likelihood; deltaL: difference in likelihood; p-AU: p-value of the approximately unbiased (AU) test.

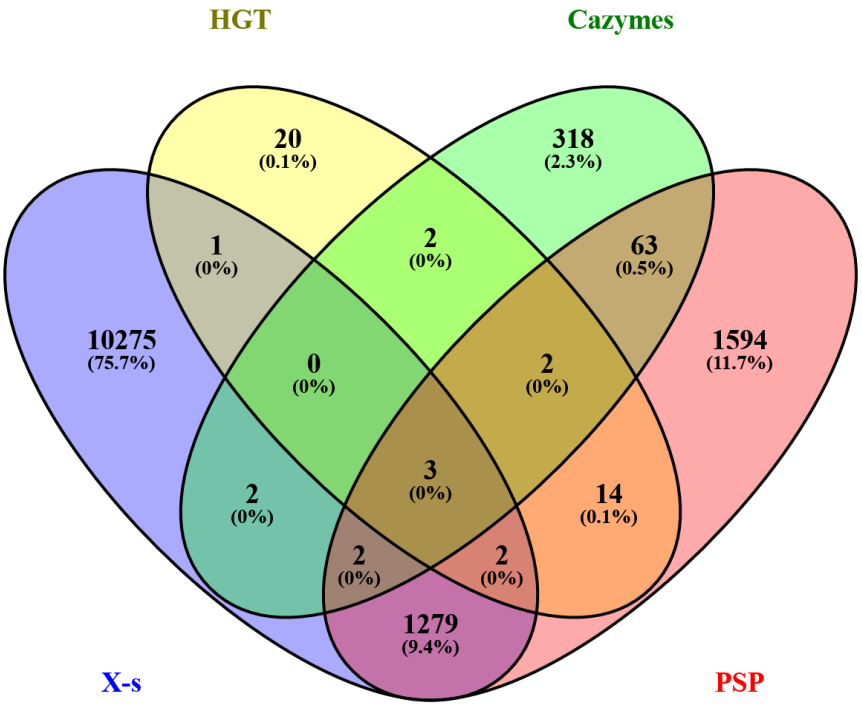



**Figure 5 :** Venn diagram between the 2959 PSP, 44 validated HGT, 11,564 X-s and 400 CAZymes. Produced with Venny 2.1.0.

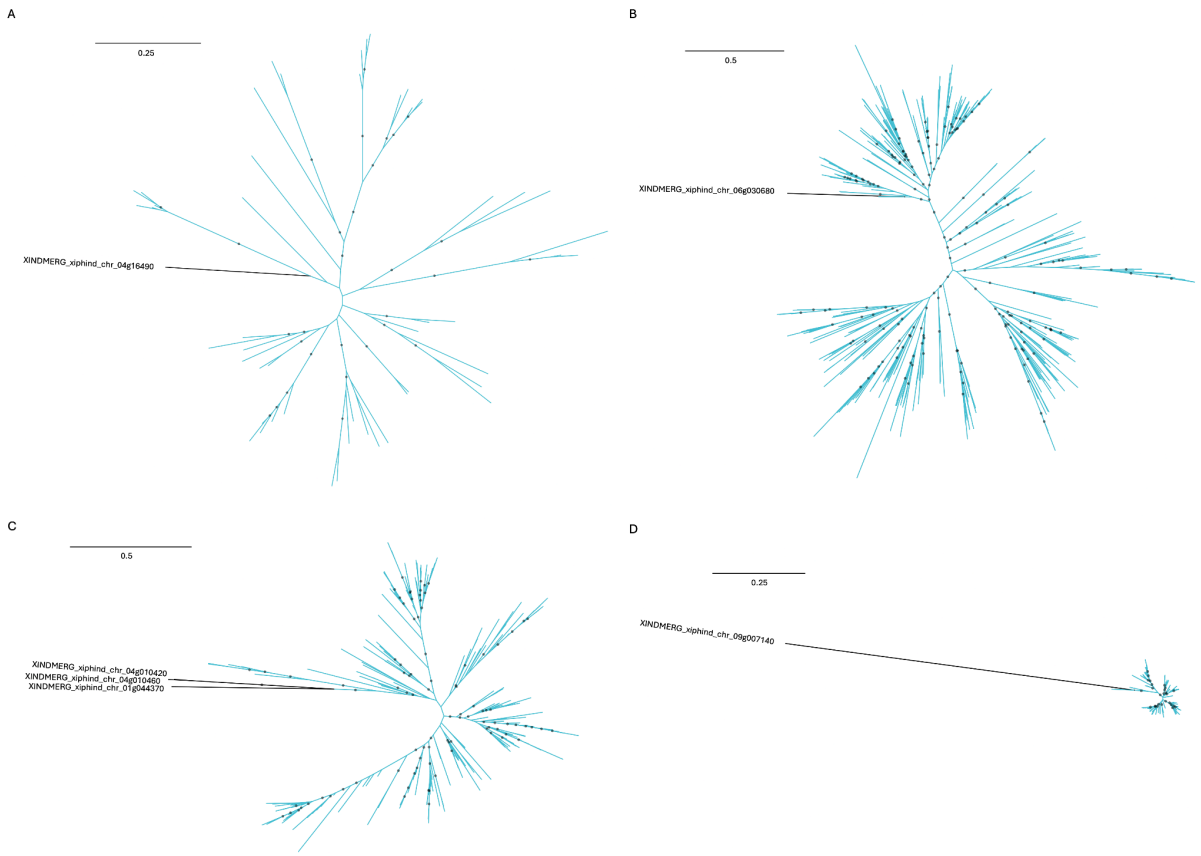



**Figure 6 : Phylogenetic analysis of the HGT corresponding to CAZymes.** The *X. index* sequence is shown in black. Blue branches represent bacterial sequences. Gray circles indicate nodes with support values greater than or equal to 80% and 90% for SH-aLRT and UFboot, respectively. (A) XindB05 HGT event corresponding to GH28. (B) XindB07 HGT event corresponding to GH32. (C) XindB12 HGT event corresponding to GH12. (D) XindB17 HGT event corresponding to GH31_2. For XindB17, the CD-HIT analysis was performed with an identity threshold of 98% since all bacterial sequences from the Diamond analysis were very similar.

# The *X. index* epigenetic toolkit is unprecedented in nematodes

## The epigenetic arsenal of *X. index* contains a complete DNA methyltransferase (DNMT) toolkit, a full-length chromatin remodeler ATRX and CTCF

We investigated the epigenetic toolkit encoded in the *X. index* genome. We first mined the *X. index* predicted proteome and the genome assembly for the presence of genes encoding DNA methyltransferases using the human (*Homo sapiens*) DNMTs as queries (hDNMT1: P26358; hDNMT2: O14717; hDNMT3: Q9UBC3) (**Figure 7**). Notably, we identified three DNMT1 candidates, from the list of *X. index*-specific proteins: xiphind_chr_02g008810 (dnmt1a), xiphind_chr_03g023520 (dnmt1b) and xiphind_chr_01g045810 (dnmt1c). As opposed to other DNMT1 described in nematodes to date [23–27], these 3 proteins contained both the canonical N-terminal domains including the foci-targeting domain (RFD) and C-terminal domains of DNMT1 (**Figure 7)**. Besides, we also found one full DNMT2

(xiphind_chr_02g036660) and one DNMT3 (xiphind_chr_02g037650) proteins in *X. index* completing the whole DNMT arsenal. Corroborating these findings, several DNA-methylation GO terms and InterPro domains were found as specific to *X. index*, as compared to other Dorylaimia nematodes (**Tables S3** and **S4**). Regarding these unprecedented results, we wondered if *X. index*'s genome holds other major epigenetic players. To date, it is generally considered that nematodes display only a truncated version of the chromatin remodeler ATRX [43], named XNP-1 in *C. elegans* [44]. XNP-1 lacks the ATRX ADD domain also present in the DNMT3. Thus, we mined the genome of *X. index* looking for a full length ATRX gene. Similarly, and in addition to the xnp-1 ortholog (xiphind_chr_02g004130), we found one gene coding for a full ATRX protein (xiphind_chr_03g035120) containing the ADD/PHD N-terminal domain. To our knowledge, besides vertebrates, *X. index* is the first animal described to have the full chromatin remodeler ATRX. Based on RNA-seq data, dnmt1a, dnmt1b, dnmt1c, dnmt2, dnmt3 and *ATRX* genes in *X. index* are all expressed to various extents (**Figure S7**).
Using a combination of BLASTp, tBLASTn, and HMM analyses, we further investigated the presence/absence of canonical *DNMTs* and *ATRX* in our dataset of 71 nematode genomes and proteomes to obtain a comprehensive representation of the distribution of these genes in the nematode tree of life. Besides the three dnmt1 genes identified in *X. index*, we only found one other candidate in *Mesodorylaimus yzb2* and confirmed the lack of evidence for this enzyme in the rest of the nematode species studied. The genomic resources described above were selected to include only highly contiguous assemblies. However, the basal clades (1 & 2) of the Nematoda phylum were under-represented due to the lack of genomic resources in these clades, thus preventing us from precisely tracing the evolutionary history of such proteins. To complement this study, we mined 60 additional draft nematode genomes produced from a recent phylogenomic analysis of the nematode phylum [13]. Doing so, we identified five other species with full-length DNMT1 and seven with partial evidence for DNMT1. Interestingly all these species belong to the Dorylaimida order. No other full-length DNMT1 was detected in the other orders (**Figure 7, Figure S8**). Conversely, the distribution of canonical DNMT3 and ATRX proteins was restricted to the basal clades 1 & 2, suggesting a loss in the last common ancestor of all the other clades. Finally, the DNMT2 was found in some Spirurina, Rhabditina and up to some basal Tylenchina, but not in Tylenchymorpha, suggesting a more recent loss (**Figure 7**). These results highlight the key phylogenetic position of the Dorylaimida in highlighting the epigenetic toolkit shift.
The unexpected presence of both canonical *DNMT1* and *ATRX* genes in Dorylaimida drove us to search for the known interactors of these well-studied proteins. Some of the interactors are found in all nematodes while others appear to be restricted to basal clades. Overall, we found that most of the known interactors of ATRX and DNMT1, including the interactant shared between these two proteins, have an ortholog in *X. index* suggesting fully conserved pathways **(Figure 7)**.

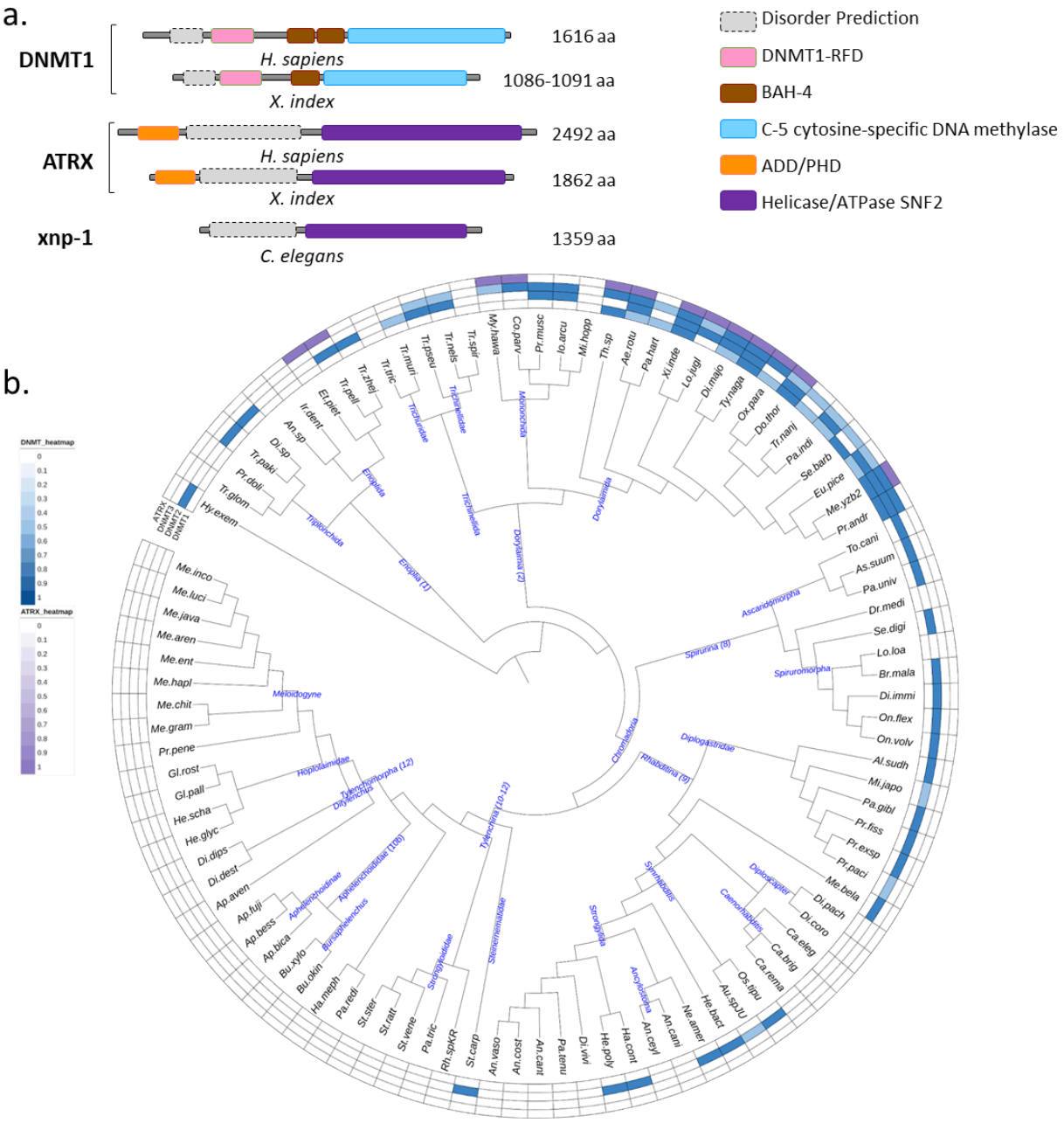

a.
DNMT1
H. sapiens
1616 aa
X. index
1086-1091 aa
ATRX
H. sapiens
2492 aa
X. index
1862 aa
xnp-1
C. elegans
1359 aa
Disorder Prediction
DNMT1-RFD
BAH-4
C-5 cytosine-specific DNA methylase
ADD/PHD
Helicase/ATPase SNF2
b.
DNMT_heatmap
ATRX_heatmap
c.
ATRX Interactors
ATRX & DNMT1 Interactors
DNMT1 Interactors
PDS5B
SMC1
SMC3
Cohesin complex
MeCP2
MBD2/3
HDAC1
HDAC2
DMAP1
DNMT3B
DNMT3A
Repressive transcription complex
FACT complex
BRD4
SSRP1
SUPT16H
Transcription repression
EZH2
ATRX
SETDB1
TRIM28
ZNF274
ZNF304
DNMT1
UHRF1
USP7
PCNA
FANCD2
HR-mediated DNA DSB repair
Heterochromatin maintenance
H3
MRE11
RAD50
NBS1
MRN complex
HP1α/CBX5
DAXX
Number of orthologs/homologs found in X. index
Found in most nematode orders
Restricted to Dorylaimida*
Not found in nematodes

**Figure 7: a.** Domain structure of the canonical ATRX and DNMT1 proteins. The illustration highlights characteristic crucial domains that define these proteins (i.e. RFD and C-5 DNA methylase, and ADD/PHD and Helicase/ATPAse SNF2 domains for DNMT1 and ATRX proteins , respectively). The binding sites localization of the interactors DMAP-1, CBX5, MECP2, EZH2 and DAXX, which have been demonstrated in humans, are indicated. **b.** Phylogenetic tree illustrating the presence/absence of canonical DNMT1, 2 and 3, and ATRX proteins in 103 nematode species (the 71 included in the orthofinder analysis and 32 draft nematode genomes from Dorylaimia and Enoplia produced in [13]. Dark blue (DNMTs) and purple (ATRX) represent the presence of the corresponding full-length protein-coding gene. Light blue indicates partial evidence for the presence. **c.** Main DNMT1 and ATRX interactors. Circles indicate the ortholog/homolog numbers found in the predicted proteome of *X. index* and colours indicate if orthologs are detected in other nematodes. H3: Histone 3, DAXX: Death domain-associated protein 6, CBX5 (HP1α): Chromobox protein homolog 5, EZH2: Enhancer of Zeste homolog 2, PCNA: proliferating cell nuclear antigen, NBS1: Nibrin, MRE11:Double-strand break repair protein MRE11; MeCP2: Methyl-CpG-binding protein 2;SMC1: Structural maintenance of chromosomes protein 1; SMC3: Structural maintenance of chromosomes protein 3; BRD4: Bromodomain-containing protein 4; SSRP1: FACT complex subunit SSRP1; SUPT16H:FACT complex subunit SPT16; FANCD2:Fanconi anemia group D2 protein; RAD50: DNA repair protein RAD50; SETDB1: Histone-lysine N-methyltransferase SETDB1; TRIM 28: Transcription intermediary factor 1-beta; ZNF274: Neurotrophin receptor-interacting factor homolog; ZNF304: Zinc finger protein 304; USP7: Ubiquitin carboxyl-terminal hydrolase 7; UHRF1: E3 ubiquitin-protein ligase UHRF1; DNMT3A/B: DNA (cytosine-5)-methyltransferase 3A/B; DMAP1: DNA methyltransferase 1-associated protein 1; HDAC1 & 2:Histone deacetylase 1 & 2; MBD2 & 3: Methyl-CpG-binding domain protein 2 & 3; PDS5B:Sister chromatid cohesion protein PDS5 homolog B; ATRX: Transcriptional regulator ATRX; DNMT1: DNA (cytosine-5)-methyltransferase 1. *including or not nematodes from basal clades (1 & 2).

Finally, we wondered if the peculiarity of the epigenetic toolkit encoded in *X. index's* genome could be extended more largely to the regulation of the genomic three-dimensional (3D) organisation. The insulator CCCTC-binding factor (CTCF), along with cohesin, play crucial roles in the 3D organisation of the mammalian genomes [45]. While members of the cohesin complex such as SMC1 and SMC3 are present in most nematode orders (**Figure 6c**), CTCF is not conserved in *C. elegans* [46]. We thus assessed for the presence of CTCF orthologs in *X. index* using human CTCF as reference and identified the gene xiphind_chr_10g011800 as *X. index CTCF* encoding gene. Based on the analysis carried out on the 71 nematode proteomes, putative CTCF were only detected in Dorylaimida and other basal orders suggesting that as DNMTs and ATRX, CTCF was part of the important reshuffling that occurred in the nematode epigenetic toolkit through evolution.

## DNA C5 methylation at CGs in *X. index* anti-correlates with gene expression

The presence of the full DNMT arsenal in *X. index* suggests DNA methylation might exist at CGs in this species. Thus, we took advantage of nanopore sequencing data to assess the presence of mCGs in the genome of *X. index* [47]. In addition, we performed an Assay for Transposase-Accessible Chromatin using Sequencing (ATAC-seq) on three biological replicates of *X. index* to assess the chromatin accessibility throughout the genome. DNA methylation within the promoter region being largely associated with gene silencing and reduced chromatin accessibility [48], we investigated the gene expression patterns, the

methylation status at CGs and the chromatin accessibility (ATAC-seq) surrounding the translation start sites genome wide. A k-means clustering based on the percentage of methylated CGs highlighted three clusters (**Figure 8**). The meta-analysis yielded a first cluster (cluster 1) containing 2932 lowly-expressed genes with a high percentage of CGs in the TSS region, a high proportion of which being methylated. In line with gene regulation observed in vertebrates [29,48], the chromatin accessibility was relatively low in this cluster (**Figure 8** and **Figure S9**). By contrast, the cluster 3 included 9467 highly expressed genes associated with low to null proportion of methylated CGs and increased chromatin accessibility across the promoter regions (**Figure 8** and **Figure S9**).
Altogether, these results confirm the presence of DNA methylation in *X. index* and suggest an important role in the chromatin landscape remodeling and gene expression orchestration such as the one described in vertebrates.

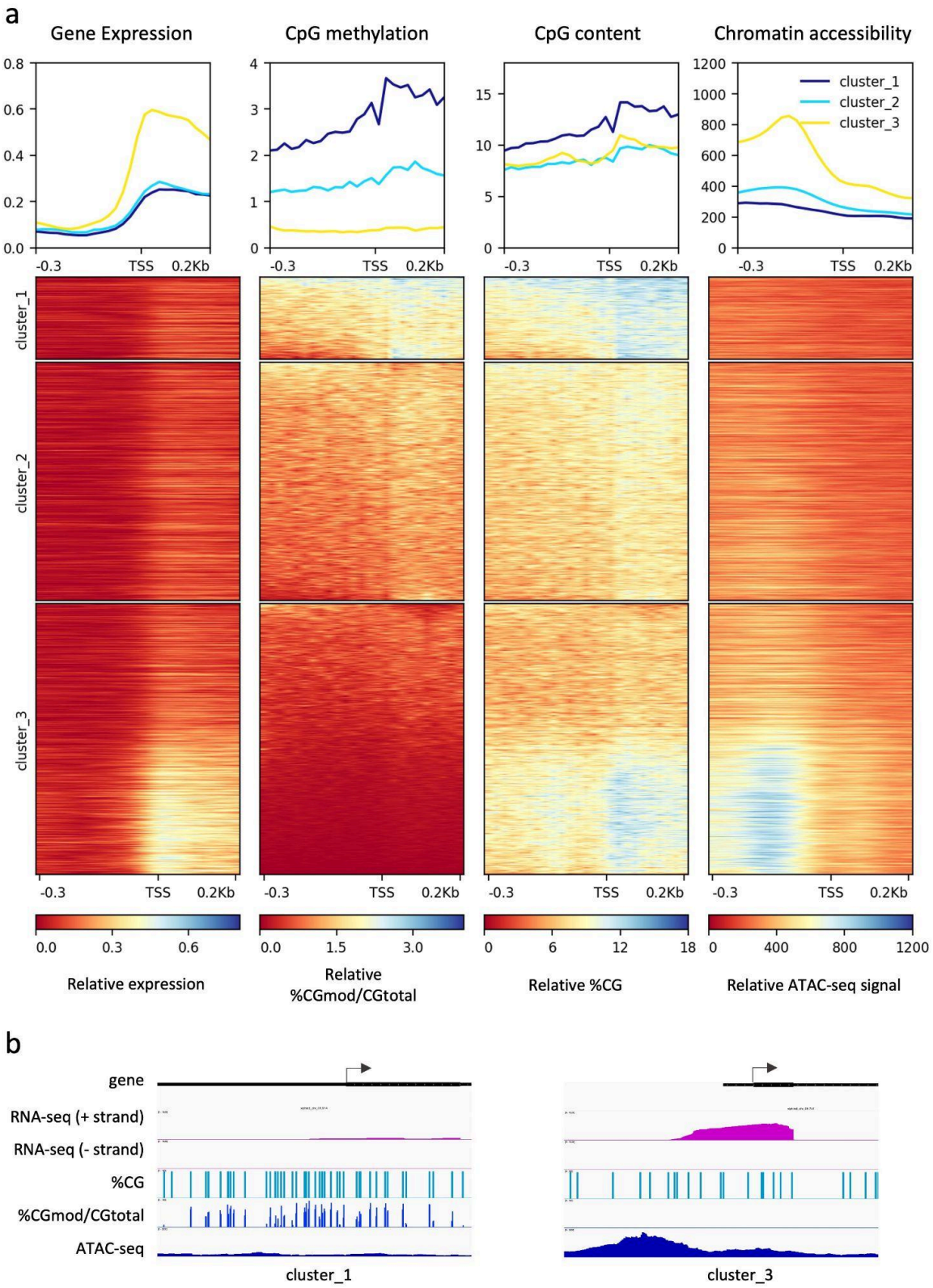



**Figure 8:** DNA C5 methylation status at CGs in *X. index* anti-correlates with gene expression and open chromatin. **a**. heatmaps and metaplots showing the relative levels of gene expression, percentage of CGs methylated (CGmod) over total CGs, %CGs and chromatin accessibility from 300 bp upstream to 200bp downstream the translation start site (TSS) of protein coding genes based on k-means clustering performed on CGmod. **b**. Examples of candidate regions (500bp windows) surrounding the TSS (arrow showing the TSS and the gene strand).

# Discussion

Here, we produced and analyzed the first telomere-to-telomere assembled nematode genome in the basal Dorylaimida order. Comparative analysis with other nematode genomes revealed a completely unparalleled epigenetic toolkit, reshuffling our knowledge on the invertebrate epigenetic machinery. Our analysis also revealed the first candidate HGT of plant origin in a nematode and several other *Xiphinema*-specific features, highlighting the interest of mining poorly covered phylogenetic branches of the tree of life. Besides genetic and evolutionary biology discoveries, this resource will be useful towards better combating the GFLV's vector, *X. index,* which is a major threat to vineyards worldwide.

## A major reshuffling of the epigenetic machinery occurred in the Nematoda phylum

Genome-wide methylation at CGs along with full repertoires of DNMTs is prevalent in vertebrates and involved in the epigenetic regulation of gene expression. However the picture in invertebrates, which represent the vast majority of animals, is much more disparate including no or mosaic CGs methylation, and frequently incomplete or absent DNMT arsenal. The presence of complete repertoires of DNMTs and consistent 5meC levels in nematodes are scarce[23,27] Although some whole clades such as Rhabditina and Tylenchida seem to have lost both the enzymes and CGs methylation, N6-adenine methylation have been highlighted in *C. elegans*[25]. In more basal clades, presence of DNMT1 has been previously reported, however our analysis detected none of DNMT1 canonical N-terminal domains in this previously reported species. This lack of conservation of canonical full-length DNMT1 suggests that methylation at CGs would not be a common mark used for gene regulation in nematodes. Yet, our analysis in *X. index,* points out (i) a full and large arsenal of DNMT, (ii) a global genome methylation at CGs and (iii) an anti-correlation between 5meC levels at TSS and gene expression and accessibility of the chromatin. This ensemble of findings suggests some unexpected similarities between the epigenetic arsenal of this nematode and vertebrates[32]. Nevertheless, why three expressed copies of the guardian of the DNA methylation maintenance (*i.e.* DNMT1) were found in *X. index* is still an open question as only one copy is found in humans for instance. In the end our analysis in other nematodes suggest that this process/toolkit might be conserved in Dorylaimida species but not in other clades, However, future studies targeting 5meC on these organisms should confirm this hypothesis.

In addition, *X. index* displays a full-length version of the chromatin remodeler ATRX, a revolution outside vertebrates. This key player is well known in genetic diseases and cancer research. The association between ATRX deficiency and aberrant DNA methylation was raised decades ago[49] and has been strengthening since then[29,50] suggesting that ATRX and DNMTs are intrinsically linked to maintain the chromatin integrity and the fine tuning of specific genomic targets.

In vertebrates, the conservation of the telomere sequence TTAGGGG plays a critical role for the binding of the shelterin proteins and the maintenance of the telomere integrity[51]. Our results revealed a very sparse conservation of the *C. elegans* sheltering complex in *X. index*, despite the presence of an identical telomeric repeat sequence in both species. Sheltering proteins are yet essential to protect chromosome ends and prevent them from being confused with double stranded DNA breaks. The presence of the full-length ATRX in *X.*

*index* and its role in the maintenance of the telomere integrity, DNA replication raises the hypothesis that it could play a key role in *X. index.* This hypothesis is further supported by the presence of ATRX interactants such as MRE11, member of the MRN complex, which play key roles in genomic stability and double strand break repairs [52]. Thus, the presence in *X. index* of canonical telomeric repeats (TTAGGC)n, a telomerase and ATRX concommitment with the absence of orthologs of the *C. elegans* sheltering complex made it a valuable model to understand the role of ATRX in the maintenance of telomere integrity and how different the telomeric machinery evolved despite a conserved telomere sequence.
Completing this epigenetic toolkit, we found a putative CTCF coding gene in *X. index*'s genome. This transcription factor/insulator is involved, along with the cohesin complex, in the 3 dimensional organisation of the genome [45]. By contrast with the cohesin complex, CTCFs were only identified in Dorylaimida and basal order as ATRX. A relationship has been described between ATRX and CTCF in mammals as ATRX deficiency alters the CTCF occupancy and so consequently chromatin interactions [53,54]. Similarly, some CTCF bindings have been shown to be methylation sensitive and thus associated with DNMTs activity, thus bringing the story full circle[55,56]. The epigenetic toolkit of *X. index* comprising DNMTs, ATRX and CTCF constitutes a regulatory network with the potential of controlling chromatin landscape and architecture and participating in the regulation of gene expression resembling vertebrates.
According to the presence / absence of canonical DNMTs, CTCF and ATRX coding genes highlighted in our study, it is clear now that a major reshuffling of epigenetic components occurred after the Dorylaimida order. More specifically, we can infer that a drastic erosion of the nematode epigenetic toolkit occurred in the last common ancestor of Chromadorea, which encompasses the Spirurina, Rhabditina and Tylenchina orders, and that can be dated ca. 482 MYA (408 - 561) according to [13]. Overall, the big picture suggests completely two different epigenetic organizations between different nematode clades. The nematodes from the Dorylaimida order share striking major similarities with the epigenetic machinery described in vertebrates, making them the most ancient animals to possess all these epigenetic key players and potentially relevant models for future studies. By contrast, members of the Chromadorea class may have developed their own epigenetic regulation or share some similarities with other invertebrate organisms as no evidence for full-length DNMT1 or ATRX are present. Future studies based on genomic resources in other early branching nematodes from Chromadorea (*e.g.* Chromadorida, Monhysterida and Plectida) would probably confirm this clear split in the phyllum.

## A first HGT from plants to nematodes

We initially identified 1,449 possible HGTs (AI or AHS>0) in the genome of *X. index,* which is consistent with numbers described in other PPNs (reviewed in [42]). Our analysis identified 44 highly likely HGTs, that mainly originated from fungi and bacteria. The detected HGTs essentially concerned enzymes, such as glycoside hydrolases from diverse families among which some have already been characterized in other PPNs (*e.g.* GH 32 in cyst nematodes or GH32 and 12 in *Longidorus*) [12,57,58]. A pioneer study based on the transcriptomes of *X. index* and *Longidorus elongatus* identified 62 putative HGTs in *X. index*, including GH32 and GH12 enzymes [12], which were confirmed by the present study. Among the HGTs identified in this study but not in the previous study, is a member of the GH28 family.
A recent and comprehensive study showed that HGT events of plant origin are extremely rare, if not absent in PPNs [42]. To our knowledge, we describe here the first likely case of

gene transfer from a plant to a nematode. This gene codes for a calmodulin protein. Many calcium-related processes in eukaryotic organisms depend on calmodulins, which act as sensors by binding to diverse signalling proteins, such as kinases. Strikingly, this plant-derived HGT is presumably secreted, since a signal peptide was detected in the nematode protein, while no such signal sequence was detected in the closest plant sequences. This makes it a compelling putative effector to investigate further in future functional studies. A recent study in rice identified a root-knot nematode effector that interacts with a calmodulin protein. Interestingly, this interaction diminishes another interaction, between this calmodulin and a conserved DNA binding protein through a competitive process. This results in the release of the latter, ultimately leading to enhanced plant susceptibility to the parasite [59]. It is plausible that the plant-derived HGT, secreted by *X. index* into plant hosts, in a sort of return to sender, competes with plant endogenous calmodulins and consequently unsettles other interactions, thus promoting the infection of *X. index*. More broadly, plant stress responses are associated with transient increases in cytosolic calcium, which can be detected by calcium-sensor proteins, such as calmodulins. Subsequently, calmodulin-dependent protein kinases, which are dually regulated by calmodulin and calcium through direct interaction, trigger downstream plant responses [60]. We could postulate that exogenous calmodulin interferes with these interactions and, ultimately, with the plant's response to the nematode. Even though these hypotheses are appealing, we can still legitimately wonder why plant-derived HGT is not more widespread in PPN. It should be noted here that other PPN are known to secrete short peptides that are highly similar to plant ones [61,62]. Due to the short length of these so-called plant mimics, homology searches and phylogenetic analyses are complicated, if not impossible. Therefore, whether these peptides have plant origin and were acquired via HGT or emerged in nematodes via convergent evolution remains unresolved.

## The heterozygote genome of *X. index* includes NUMTs

*X. index* is a diploid parthenogenetic species with occasional sexual reproduction [63], that shows an estimated heterozygosity level of 0.9%. In plant-parasitic nematodes, obligatory parthenogenetic species with dysfunctional meiosis harbor significantly higher heterozygosity levels (7-8%), which is explained by their hybrid origin and polyploidy [35,64–66]. In the nematode *Mesorhabditis belari* from the Rhabditida order, the level of heterozygosity, obtained from 10 different strains, is slightly higher, about 1.3%. This diploid species reproduces by parthenogenesis involving the fusion of the two products of meiosis I (central fusion of homologous chromosomes) [67]. The reduction of heterozygosity is slowed under central fusion whereas a faster reduction of heterozygosity would be observed in terminal fusion which results in the unification of sister chromatids after meiosis II [68]. Although the relatively high heterozygosity (0.9%) of *X. index,* might suggest central fusion of non-sister chromatids, the exact process involved, i.e. central or terminal fusion, still needs to be experimentally elucidated at the cytological level.

The last common ancestor of eukaryotes possessed a mitogenome containing presumably 69 protein-coding genes. Many of these genes have been lost or transferred to the nuclear genome, via endosymbiont gene transfer (EGT), in different proportions depending on lineages [69]. The mitochondrial genome of *X. index* contains a set of 12 protein coding genes that corresponds to the core set of mitochondrial proteins in most other nematodes (except Trichinellida species that additionally possess the *ATP8* gene). The mitogenome also contains some non-genic regions like some other nematodes. The evolutionary origin and

role of these non-coding sequences, which lack homology in other species, remain unknown. Similarly, there is no evident explanation for their disparate persistence in *X. index* and some species while other species have very optimized and gene-dense mitogenomes (*i.e.* the full sequence is dedicated to code for rRNA subunits, tRNA or protein coding genes like in *Xiphinema* species of the american group). NUMTs are now known to occur in many lineages, and such transfer is an ongoing process, which can be considered as a type of intracellular HGT [70]. In the insect order Orthoptera, a large number of NUMT have been identified in many species, even revealing a correlation between number of NUMT and nuclear genome size. These insertions, of *a priori* non-functional fragments of mtDNA, were classified as recent insertion (neonumts) or ancient (paleonumts). Some paleonumts originated tens of millions years ago which suggest that these “non-functional” sequences can remain present for a long period [71]. In our analysis, we used stringent parameters, which probably underestimates NUMTs number, and chose to focus on the largest one by far (K10 insertion). We showed that the full mitochondrial genome has been transferred within the nuclear genome through several insertions, the K10 insertion alone counting for approx. 80% of the full mitogenome. This latter NUMT is ancient as we detected it in another unrelated population of *X. index* originated from the putative cradle of the species (i.e. the middle east) [18]. Additionally, this K10 insertion may be used for population identification. Future genomic resources in Dorylaimida species would help to date these NUMTs, thus contributing to deciphering evolutionary history of Dorylaimida species. Although NUMTs may be deleterious depending on the location of the insertion and a problem for DNA barcoding-based studies, they represent a significant source of genomic variation.

# Methods

## Biological material preparation and sequencings

### Nematode production and sample preparation

The *Xiphinema index* population ‘samos’ and ‘Iran’ were reared on fig trees in the ISA greenhouse facility, which was maintained between 5-32°C for decades and regularly verified using molecular markers[72,73] Nematodes were extracted from soil using Oostenbrink elutriator according to the protocol defined by Hooper[74]. For all the experiments related to genome sequencing and assembly, *X. index* individuals, at different stages, were picked up by hand to avoid contamination with other nematodes. 30 000, 20 000 and 6 000 *X. index* were collected in 1X antimycotic solution (Sigma) for the PACBIO HiFi, Hi-C and Novaseq sequencings, respectively. For the transcriptome, 2 000 *X. index* were soaked either in tap water, in 0.2 M NaCl for 1h or in 0.1% acetic acid for 1 h. The total of 6 000 nematodes were pooled in one tube before RNA extraction. The mix of control and stress individuals was made to maximize the number of expressed genes. For ATAC-seq experiments, 1 500 mixed stages *X. index* per replicate were collected after overnight filtration on tissues and washed with 1X PBS. For all the samples, the liquid was removed using a syringe and microtubes were flash frozen in liquid nitrogen and stored at -80°C before nucleic acid extractions.

## High molecular weight DNA extraction and library preparation for long read sequencings

The DNA extraction for subsequent long read sequencing was performed using the MasterPure™ Complete DNA and RNA Purification Kit from epicentre. Approximately 30 000 *X. index* individuals were used for HMW DNA extraction. The sample was directly ground in liquid nitrogen using a micro pestle directly in the microtube. 300 µl of tissue and cell lysis solution and 1 µl of proteinase K were added and homogenized. All homogenisation steps were performed gently to avoid DNA fragmentation. The tube was incubated at 65°C for 15 minutes, then cooled down at 37°C before adding 1 µl of 5 µg/µl RNase A and finally incubated for 30 minutes at 37°C. The sample was let on ice for 5 minutes and 175 µl of MPC protein precipitation reagent was added to the sample and mixed. The debris were pelleted by centrifugation at 4°C for 10 minutes at ≥10,000 x g. The supernatant was transferred to a clean microcentrifuge tube. 500 µl of isopropanol was added and the tube was inverted several times before centrifugation at 4°C for 10 minutes. Isopropanol was removed and the pellet was rinsed 2 times with 70% ethanol and let dry before solubilization with 120 µL of EB buffer (Qiagen). DNA was analyzed for quality and quantity controls using nanodrop, Qubit and fragment analyser. The final sample used for library preparation contained 16.4 µg of gDNA with an average fragment size of 36 789 bp. Library preparation and sequencing were performed according to the manufacturer's instructions "Procedure & Checklist Preparing HiFi SMRTbell Libraries using SMRTbell Express Template Prep Kit 2.0". At each step, DNA was quantified using the Qubit dsDNA HS Assay Kit (Life Technologies). DNA purity was tested using the nanodrop (Thermofisher) and size distribution and degradation assessed using the Femto pulse Genomic DNA 165 kb Kit (Agilent). Purification steps were performed using AMPure PB beads (PacBio). 15µg of DNA was double purified then sheared at 20kb using the Megaruptor3 system (Diagenode). Using SMRTbell Express Template prep kit 2.0 on 5µg of sample, the library was treated with an exonuclease cocktail to digest unligated DNA fragments. A size selection step using a 8kb cutoff was performed on the BluePippin Size Selection system (Sage Science) with "0.75% DF Marker S15-20 kb Improved Recovery" protocol. Using Binding kit 2.0 kit and sequencing kit 2.0, the primer V2 annealed and polymerase 2.0 bounded library was sequenced by diffusion loading onto 1 SMRTcell on Sequel2 instrument at 80pM with a 2 hours pre-extension and a 30 hours movie. Regarding the ONT sequencing, library preparation and sequencing were performed according to the manufacturer's instructions "1D gDNA selecting for long reads (SQK-LSK109)". At each step, DNA was quantified using the Qubit dsDNA HS Assay Kit (Life Technologies). DNA purity was tested using the nanodrop (Thermofisher) and size distribution and degradation assessed using the Fragment analyzer (Agilent) DNF-464 HS Large Fragment Kit. Purification steps were performed using AMPure XP beads (Beckman Coulter). For 1 Flowcell, 3µg of DNA was purified then sheared at 25kb using the megaruptor system (diagenode). Using SQK-LSK 109 kit (ONT) on 2µg of sample, library was loaded onto 1 R9.4.1 flowcell and sequenced on GridION instrument at 30 fmol within 48H. Basecalling was done in super-high accuracy mode with Dorado.

## Extraction of nuclei, library preparation and illumina sequencing for Hi-C sequencing

Approximately 20 000 *X. index* were resuspended in a dounce tissue grinder tube with 2 ml of an extraction buffer containing KCl, NaCl, MgCl2, EDTA, Tris/HCl, Protease inhibitor,

sucrose, sodium butyrate, PMSF, DTT, and IGEPAL (at the final concentrations 60 mM, 15 mM, 5 mM, 0.1 mM, 15 mM, 1 g/L, 300 mM, 5 mM, 0.1 mM, 0.5 mM and 0.4%, respectively). Nematodes are crushed for 10 minutes on ice.Then, leave for 5 minutes on ice. Still on ice, the resulting solution was retrieved and slowly load in a Corex centrifugation tube containing 8 mL of a separation buffer of made of KCl, NaCl, MgCl2, EDTA, Tris/HCl, Protease inhibitor, sucrose, sodium butyrate, PMSF, and DTT (at the final concentrations 60 mM, 15 mM, 5 mM, 0.1 mM, 15 mM, 1 g/L, 1.2 M, 5 mM, 0.1 mM, and 0.5 mM, respectively). Centrifuge at 10 000 g for 30 minutes at 4°C with slow deceleration. The supernatant is removed entirely. The tubes were removed from ice and the pellet was carefully resuspended by pipetting in 5 mL of PBS 1X. Then, 500 µL of freshly prepared buffer TC containing NaCl, 100 mM; EDTA-pH8, 1 mM; EGTA-pH8, 0.5 mM; HEPES-pH8, 50 mM and Formaldehyde 22 %, was added.The solution was mixed by pipetting and let for 7 minutes at room temperature. From this step, the samples were treated using the reagents of the Arima - Hi-C kit. 289 µL of the Stop Solution 1 (from the kit) was added and the tube was incubated at RT for 5 minutes. The tube was centrifuged at 2 000 g for 15 min at RT and the supernatant was removed. At this step an aliquot of 10% of the initial volume was prepared and stored at -80°C for future estimation of the initial quantity of nuclei.Then, the pellet was resuspended in 24 µL of conditioning solution and incubated at 62°C for 10 minutes. The final steps of the protocol followed the instructions provided by the Arima-Hi-C kit (ref. 510008) along with the TruSeq DNA PCR-Free Kit and TruSeq DNA UD Indexes (Illumina, ref. 20015962, ref. 20020590) and the KAPA library amplification kit (Roche, ref. KK2620). The Hi-C library was then sequenced in paired-end 2x150nt mode on a NovaSeq6000 (Illumina).

## RNA extraction, library preparation and illumina sequencing for transcriptome

A total of 6 000 *X. index* (including untreated and stressed nematodes), were used for RNA extraction. The samples were directly ground in liquid nitrogen using a micro pestle directly in the microtube. 800 µl of extraction buffer (CTAB 2.5%, PVPP 2%, Tris-HCL 100mM, EDTA 25 mM, NaCl 2M, β-mercaptoethanol 2%) was added to the tube. After 30 minutes of incubation at 65°C, 800 µl of Chloroform/Isoamyl alcohol (CI; v/v; 24/1) was added, and after homogenization, the tube was centrifuged (16000g - 8 min - 4°C). The supernatant was retrieved and the same volume of water-saturated Phenol (pH 4.5-5) / Chloroform / Isoamyl alcohol (PCI; v/v; 25/24/1) (ca 700 µl) was added and centrifuged. A second step with CI was carried out and the supernatant was retrieved and mixed with 500 µl NaCl 5M and 500 µl of isopropanol and stored overnight at -20°C. After centrifugation (16000g - 20 min - 4°C), two cleaning steps were carried out with ethanol 70%. The pellet was let dry until the absence of ethanol residue and resuspended in 30 µl of buffer EB (Qiagen). Genomic DNA was removed from the sample using the kit TURBO DNA-free (Ambion) following the supplier's instructions. Samples purity and quality were assessed with a nanodrop and Qubit before library preparation. RNA-seq libraries have been prepared according to Illumina's protocols using the Illumina TruSeq Stranded mRNA sample prep kit to analyze mRNA. Briefly, mRNA were selected using poly-T beads. Then, RNAs were fragmented to generate double stranded cDNA and adaptors were ligated to be sequenced. 11 cycles of PCR were applied to amplify libraries. Library quality was assessed using a Fragment Analyser and libraries were quantified by QPCR using the Kapa Library Quantification Kit. RNA-seq experiments have been performed on an Illumina NovaSeq. 6000 using a paired-end read length of 2 × 150 pb with the Illumina NovaSeq. 6000 sequencing kits.

## Nuclear genome assembly and scaffolding

First, Nanopore reads were filtered for a minimum quality of Q10 and a minimum length of 30kb using Nanofilt 2.8.0 [75]. Then, we used HiFiasm[76] version 0.19.8 with the PacBio HiFi long reads combined with the Nanopore longest reads (-ul option) and the Hi-C Illumina paired-end short reads (-h1, -h2 options) with the other parameters left to default to produce two haplotype assemblies and a merged consensus assembly (primary contigs). We used TiDK [77] to search the canonical nematode telomeric repeat TTAGGC in the contigs. Then Juicer (v 1.6) from 3D-DNA [78] combined Hi-C data and each assembly to produce initial contact maps and scaffolding. For each assembly, final scaffolding was manually edited using JuiceBox (v2.13.7) with the help of nematode telomere motifs identified using TIDK (converted in wig files). Scaffolding of the HiFi contigs with HiC data introduced gaps, so we attempted to fill the gaps with the ONT data. For that, we used tgs-gapcloser v1.2.1[79] without error correction and all the other parameters left to default. This tool resolved 51 % of all the gaps in the chromosomes. To identify and remove debris that could represent duplicated information already present in the pseudo-chromosomes, we used minimap2 v2.24 [80] to map the debris against all the chromosomes. Then, we calculated the debris coverage using bedtools v2.30.0 genomecov [81] and removed the debris that aligned on more than 90% of their length on pseudo-chromosomes (**Figure S10**). We used BlobTools [82,83] to identify and remove mitochondrial scaffolds as well as possible contaminations. Because there is a lack of closely related nematode sequence data in public libraries, two sources of evidence were used as hit files. A diamond blastx search against the NCBI's nr library as a first source and a NCBI blastn search against the nt library as a second source, both with an e-value threshold of 1e-25. Finally, pseudo-chromosomes were ordered by size and orientated according to telomere repeats, starting with reverse motif GCCTAA and ending with forward motif TTAGGC. The merged assembly was first ordered and orientated. The hap1 and 2 assembly scaffolds were named according to their correspondence to pseudo-chromosomes in the merged assembly.

## Assembly quality check

Assembly quality assessment was followed with N50 metrics, KAT, BlobTools, BUSCO (v5.3.1, metazoa_odb10, nematoda_odb10) for metazoa and nematoda (Benchmarking Universal Single-Copy Orthologs), DGENIES and IGV visualisation of minimap2 reads / contigs / scaffolds alignments.

For blobtools, the hits file was done with the diamond (v 2.0.8) blastx, the database was NR (2025-02-01) with the synchronized taxonomy from NCBI. The rest of the options were -k 5 -e 1e-25 -b 6 -c 1 --sensitive -F 15 --range-culling. The options range culling was because we have multiple genes on the chromosome scale genome, and diamond keeps the 5 best hits from a window in the scaffolds so we have a diversity of read mapped all along the chromosomes. The coverage file was done with minimap2 (v 2.24) against the pacbio reads and the option -x map-hifi. Those results were combined with blobtoolskit (v4.4.0), the added diamond result was done with the option --hit-count 10000 to take all the data together and the taxonomy was taken from NCBI at 2025-02-18.

## Genome size, ploidy and heterozygosity estimation

We used JellyFish version 2.3.0 [84] with a k-mer size of 21 to enumerate k-mers on the *X. index* HiFi reads. We generated histograms using the jellyfish histo command and adjusted the -h parameter to allow k-mer repeated up to 1 million times. To infer the ploidy level, we used SmudgePlot (v.0.2.5) [85] with default parameters. Finally, to estimate genome size and heterozygosity, we used GenomeScope2 [85] with the previously obtained ploidy levels and k-mer histograms.

## Mitochondrial genome assembly and NUMT

The mitochondrial genome was assembled using mitochodriAL circulAr DNA ReconstitutioN (https://github.com/GDKO/aladin; ALADIN v1.1) with -d N -m M options.The seed used for the assembly was the NADH dehydrogenase subunit 4 (ND4) partial gene sequence that was available from the same *X. index* population (NCBI accession number: LT996799). The assembly was performed using PacBio HiFi reads and verifications were conducted by aligning ONT reads.
NUMTs were detected using a blastn approach between the mitogenome and the nuclear genome assembly and results were filtered by alignment size (>500 bp) and percentage of identity (>90). The main NUMT detected in chromosome 10 was validated using specific primers. Two sets of primers were designed to flank the 5' and the 3' regions of the insertion: RWI5'_F2: CAGTCTCTTTGGCCTGGTCT / RWI5'_R2: CGGGGATCCTTTACTGTTTCA and RWI3'_F1: TCCAAAACATGATCCGCTCT / RWI3'_R1: GTTGGAATCCTGACCTGGGT. MyTaq™ red mix kit was used following the supplier's instructions. Genomic DNA of two different populations of *X. index* ('Samos', which is the one we have sequenced and 'Iran') were used to perform PCR (94°C-5min; 35 cycles of 94°C-30s, 58°C-30s, 72°C-1min; 72°C-5min) and gave amplicons of 470bp and 413bp for *X. index* Iran' and 494bp and 425bp for *X. index* 'Samos'. PCR products were sequenced and aligned to the genome sequence for final validation.

## Annotation of transposable elements, telomeres and other repeats

Earl Grey version 4.2.4 [86] was used to detect transposable elements (TEs) and other repeats both for the two separated haplotypes and the merged reference assemblies. Default parameters were used with the exception of the -m option and "RepeatMasker search term" (-r) nematode. These options were used to remove TEs with a size lower than 100bp and to mask repeats and low complexity DNA sequences of known nematode elements, respectively.

Telomeres were identified with TIDK (v0.1.5) using the canonical telomeric sequence for nematodes 'TTAGGC' and a window of detection of 1kb.

## Gene prediction & transcriptome assembly

Gene models prediction was done with the fully automated pipeline EuGene-EP (v2.0.3) [87]. EuGene has been configured to integrate similarities with known proteins of Caenorhabditis

elegans (PRJNA13758) from WormBase Parasite [88] and “nematoda” section of UniProtKB/Swiss-Prot library [89], with the prior exclusion of proteins that were similar to those present in RepBase [90]. The dataset of Xiphinema index transcribed sequences generated in this study were aligned on the genome and used by EuGene as transcription evidence. For this, we first assembled de novo using Trinity [91] the transcriptome of *X. index* obtained from the one above-described condition and for a given trinity locus we only retained the transcript returning the longest ORF. Finally, only de novo assembled transcripts that aligned on the genome on at least 30% of their length with at least 97% identity were retained.The EuGene default configuration was edited to set the “preserve” parameter to 1 for all datasets, the “gmap_intron_filter” parameter to 1 and the minimum intron length to 35 bp. Finally, the Nematodes-specific Weight Array Method matrices were used to score the splice sites (available at this URL: http://eugene.toulouse.inrae.fr/WAM/).

## Transcriptional support of gene models by RNA-seq reads and RNA-seq visualisation

To determine the expression of protein-coding genes, we used RNA-seq data mapped to the genome with STAR (Version 2.7.8a) using more stringent end-to-end parameters [92].
Expected read counts, FPKM and TPM values were calculated on the predicted genes from the *Xiphinema index* GFF annotation using RSEMversion 1.3.1 [93]. We considered as supported by transcription data any gene with an FPKM equal or higher than the value of the first quartile of the FPKM distribution for the whole set of predicted genes.
BigWig (bw) files mapped RNA-seq data were generated with a data normalization by Counts Per Million mapped reads (CPM) using Deeptools (version 3.5.6) bamCoverage with the options -bs 1 --normalizeUsing CPM --ignoreDuplicates. Similarly strand specific bw files were generated with the extra option --filterRNAstrand forward or --filterRNAstrand reverse. Data were visualized using IGV 2.16.2.

## Comparative analysis with other nematode genomes

To compare genome and gene content completeness as well as gene families and associated functions with other nematode genomes, we selected a dataset of other nematode species belonging to the Dorylaimia clade. From the WormBase ParaSite [88] resource, we selected Dorylaimia genomes (Clade I in Wormbase ParaSite) with a minimum of 70% BUSCO [94] nematoda completeness at the genome level according to WormBase. Nematodes from three genera fulfilled this criterion: *Trichuris*, *Trichinella* and *Mesodorylaimus*. For each genus, we selected up to two species that presented the highest N50 at the genome assembly level. This ended up with a selection of five species: *Trichinella spira*[37]*s* [37], *Trichinella nels*[38]*i* [38], *Trichuris mu*[39]*s* [39], *Trichuris trichi*[39]*a* [39], and *Mesodorylaimus sp. YZB*[40]*4* [40] as the sole available for this genus.

### BUSCO completeness

We used BUSCO v5.7.1 with the metazoa.odb10 dataset and default parameters to estimate completeness both at the genome assembly level (--genome) and on the set of predicted proteins (--proteome) for *X. index* and the above-mentioned species.

### OMArk completeness

We first used omamer (v2.0.3) search with LUCA dataset. Then we used the result to OMark (v0.3.0) with the same dataset, with -r phylum -t 46003 option.

We used OMArk v0.3.0 to evaluate the completeness of the predicted proteome of *X. index* as compared to the most closely related set of species in the omamer (v2.0.3) LUCA dataset, using default parameters. We use for omark with the taxonomy rank of phylum. Because we realized *Mesodorylaimus sp. YZB2_4* was not present in the omamer dataset despite being more closely related than the other Dorylaimia, we also ran an OMArk completeness evaluation on this species as well.

### InterPro and Gene Ontology annotation

We used InterProScan v5.68-100.0 [95] to detect conserved domains in the predicted protein set of *X. index* as well as those of the 5 other Dorylaimia species mentioned above. We used the options --goterms and -pa to allow assignment of corresponding Gene Ontology (GO) terms as well as prediction of corresponding MetaCyc and Reactome pathways.

### Orthology search

We used OrthoFinder v2.5.4 [96] to identify homologs of *X. index* predicted proteins in other nematode predicted proteomes. We started from a previous OrthoFinder analysis that included 68 nematode predicted proteomes and one tardigrade proteome as an outgr[35]p [35]. We added the *X. index* proteome as well as those of *Trichuris trichiura*, and *Mesodorylaimus sp. YZB2_4* that were not previously included. We also removed the *Trichuris suis* predicted proteome as it did not satisfy our 70% metazona BUSCO completeness minimal threshold. We used the option -M msa to generate multiple-sequence alignments and phylogenetic trees for all orthogroups.
By analyzing the produced orthogroups we then identified *X. index* proteins conserved with other nematodes as well as those specific to this species.

## Detection of DNMTs and ATRX and their interactors in nematodes

### Identification of candidate DNMT1 proteins and phylogenetic analysis

We used complementary proteome and genome-based strategies to identify candidate DNMT1 genes and proteins in the 71 nematoda species analyzed. First, we used HMMER to search the PANTHER domain PTHR10629 in the 71 nematoda predicted proteomes used for OthoFinder with an e-value threshold of 1e-80. Second, we checked whether orthologs of the candidate DNMT1 identified by HMM were present in other nematodes using the orthogroups generated by OrthoFinder. Third, we used the nematode protein with the lowest e-value returned by HMM as a query to search using tBLASTn the genome sequences of these 71 nematode species. This last step was to make sure the absence of identification at the protein level was not simply due to lack of prediction of the gene model in the genome.

To further extend the nematode dataset of candidate DNMT1 proteins, we repeated this same tBLASTn search in the 60 draft genomes that cover 11 orders, recently produced as part of a comprehensive phylogenomics analysis of the phylum nemat[13]a [13]. The tBLASTn match regions were extracted and we used AUGUSTUS and FGENSEH+ [97] with the query protein to predict genes and the corresponding amino acid sequences at the corresponding loci. The amino acid sequences were then analysed with interProScan to assess the tBLASTn results.
In the end, we complemented the collection of candidate Ecdysozoa DNMT1 proteins identified[23]n [23] by the newly identified ones. For consistency of inclusion criteria we only kept those proteins which had a match with the PTHR10629 PANTHER domain with an e-value threshold of 1e-80.
Orthologs of ATRX and DNMT1 interactors in *X. index* were searched using reciprocal blastP analysis with the human sequences as queries. Then, the *X. index* ortholog sequences identified were used as queries in BlastP analysis against ClusteredNR database restricted or not to Nematoda (taxid:6231).

### Chromatin accessibility assay and data analysis

ATAC-seq experiments were performed in triplicates using ATAC-Seq Kit from Active Motif (53150) following manufacturer's instructions. Briefly nuclei from 1 500 mixed stages *X. index* were extracted with Douncer in ATAC lysis buffer before being filtered on 40 µm followed by 10 µm mesh strainers. 200 000 nuclei were tagmented in Tagmantation Master Mix for 30 min at 37°C at 800 rpm. Fragmented DNA was purified using DNA purification columns. PCR amplification and indexing was performed using i7 and i5 Indexed Primers. Samples were then purified with SPRI Beads. Library quality was assessed using a Fragment Analyser and sequenced using a paired-end read length of 2 × 75 pb with AVITY chemistry. ATAC-seq data were trimmed with fastp (v0.23.4) and mapped using Bowtie2 (v2.5.1). Bam files were generated with samtools (v1.20). bigwig files were generated using bamCoverage from deeptools (deeptools_3.5.6--pyhdfd78af_0) with the options -bs 1 -p 4 --normalizeUsing RPKM --ignoreDuplicates and visualized on IGV 2.16.2.

## Horizontal gene transfers

### Detection of HGT candidates

A homology search was performed for the *X. index* proteome against the NCBI non-redundant (NR) protein database using DIAMOND (v2) [98] in the more sensitive mode, with an e-value threshold of $1.0e^{-3}$ and a maximum number of hits of 500. The DIAMOND homology search results were submitted to AvP [99] to calculate the Alien Index (AI) and the Aggregate Hit Score (AHS) for each query sequence. The AI and AHS metrics are based on the single best animal (NCBI:txid33208) and non-animal hits and the normalized sum of the scores of the best animal (NCBI:txid33208) and non-animal hits, respectively. Proteins with an AI or AHS greater than 0 have a higher similarity to non-animal than to animal hits in the NR database. Self-hits to Longidoridae (NCBI:txid46001) were ignored.
AvP was further used for automatic phylogenetic detection of HGT candidates among proteins with AI or AHS above 0. The first step of AvP was to cluster the query sequences based on the percentage of shared hits in the DIAMOND homology search result using a 50% overlap threshold (instead of the default value of 70%). Protein sequences with

significant hits in the DIAMOND homology search result were then retrieved from the NR database by AvP. If no animal sequences were found in the Diamond result, only protein sequences corresponding to the 50 best hits were retrieved. If animal sequences were found in the Diamond result, all protein sequences corresponding to hits preceding the first animal sequence, along with the sequences of the next 20 best hits, were retrieved.
The second step of AvP was to align sequences using MAFFT (v7) with the --auto option [100], to infer the phylogeny for each group using FastTree (v2) with the default parameters [101], and to detect HGT candidates according to the species found in the sister branch of the query sequence and the ancestral sister branch. In this initial step, FastTree (v2) was preferred over IQ-TREE (v2) [102] for phylogeny inference to improve the speed of the analysis. In a third step, AvP classified the HGT candidates according to their putative origin. The fourth step of AvP was to infer a constrained topology in which the query sequence(s) and the other animal sequence(s), if present, form a single monophyletic group and to determine whether the topology supporting HGT is significantly more likely than the constrained alternative one using an approximately unbiased (AU) test [103]. Finally, AvP analyzed the genomic environment of each HGT candidate gene and calculated a local score ranging from -1 to +1, with -1 for a HGT candidate gene surrounded by other HGT candidate genes (indicating a possible contamination, duplications after an HGT event, or multiple HGTs) or +1 for a HGT candidate gene surrounded by genes that are likely to be vertically inherited (suggesting that the contamination hypothesis can be ruled out). The local score was not reported if the number of genes in the scaffold (including the HGT candidate) was less than 5.

## Validation of HGT candidates

A manual analysis of the AvP results was performed, and HGT candidates were not further considered if at least one of the following criteria was met:
The total number of donor sequences was less than 3 in the sister branch plus the ancestral sister branch (if present). The HGT-supporting topology was not significantly more likely than the constrained alternative topology in which the query was forced to group with all animal sequences from NR. However, HGT candidates were retained if these animal sequences were either (i) taxonomically misannotated, likely due to contamination by a donor species or (ii) suspected to have originated from an HGT event themselves based on BLASTP results performed at NCBI against NR (https://www.ncbi.nlm.nih.gov/). At least one donor sequence in the DIAMOND homology search results shared more than 70% identity with the HGT candidate sequence, suggesting contamination by a donor species. The identity between the donor sequences and the HGT candidate sequence in the DIAMOND homology search results was less than 30%. Searching for homologous proteins and building a phylogenetic tree below this value is problematic due to the low quality of the alignments. The average alignment length in the DIAMOND homology search results was less than 100 amino acids with an average query coverage of less than 50%. The local score calculated by AvP from the genomic environment of the HGT candidate was less than 0, and there was no indication of duplication after an HGT event or of multiple HGTs, indicating a possible contamination.
A molecular marker named HGT1 was designed in the first intron of the gene chr_06g025050 (putative HGT from plants). We amplified HGT1 using the following primers HGT1-F: GCCACGTAAAGCTCAGCAAT and HGT1-R: GAGTCCAGACCCAAATCCCA with a classical PCR program (tm=59°C; 35 cycles) and MyTaq™ red mix kit. The PCR product size was 257bp long. Genomic DNA of two different populations of *X. index* (‘Samos’, which is the one we have sequenced and ‘Iran’).

### Phylogeny reconstruction for validated HGT candidates

For each HGT candidate detected by AvP that passed all the above criteria, we reconstructed a maximum likelihood phylogeny using IQ-TREE (v2) [102] to improve accuracy. For each group, AvP was used to retrieve the protein sequences corresponding to all 500 Diamond hits. CD-HIT analysis was then performed with an identity threshold of 90% to remove redundancy[104]. Sequences were aligned using MAFFT (v7) with the --auto option [100]. Highly divergent sequences were removed from the alignments using trimal (v1.4) with the -resoverlap 0.8 and -seqoverlap 75 options [105]. Alignments from which all *X. index* sequences were removed were rejected. For each remaining alignment, poorly aligned regions were removed using trimal (v1.4) with the -automated1 option. Alignments with less than 100 sites remaining were rejected, as the signal may not be sufficient to infer phylogenies with high confidence. For the remaining alignments, duplicate sequences were removed using SeqKit (v2) with the rmdup option [106], and phylogenies were inferred using IQ-TREE (v2) with automated model selection as implemented in ModelFinder [107]. Support values were based on a Shimodaira-Hasegawa approximate likelihood ratio test (SH-aLRT) [108] combined with an ultrafast bootstrap (Ufboot2) approximation with 1,000 replicates [109]. Only support values greater than or equal to 80% and 90% for SH-aLRT and UFboot, respectively, were considered. Phylogenies were visualized with iTol [110].
When Metaozan sequences, other than from *X. index*, were present in the phylogenetic tree, we forced a constrained topology supporting monophyly of animal sequences, and determined whether the unconstrained topology was significantly more likely than the constrained alternative one using an approximately unbiased (AU) test [111] with IQ-TREE (v2).

## Detection of encoded Carbohydrate-active enzymes (CAZymes)

All the predicted proteins from the *X. index*, *Mesodorylaimus sp. YZB2_4*, *Trichinella spiralis*, *Trichinella nelsoni*, *Trichuris muris*, and *Trichuris trichiura* genomes were compared with entries in the CAZy database [112]. A homemade pipeline combining the BlastP (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and HMMER3 (http://hmmer.janelia.org/) tools was used to compare protein models with the sequences of the CAZy modules. Proteins with E-values less than 0.1 were further screened by a combination of BlastP searches against libraries generated from the sequences of the catalytic and non-catalytic modules. HMMER3 was used to search against a collection of custom Hidden Markov Model (HMM) profiles constructed for each CAZy family. Expert curators then performed manual inspection to resolve borderline cases.

## Methylome analysis

The detection of methylated CpGs was performed by basecalling the nanopore long-read sequencing data with the option --modified-bases 5mCG_5hmCG and aligning the reads with dorado v0.9.0 on the merged assembly. A bedMethyle file was generated using Modkit v0.4.3 pileup with the option --cpg.
The data regarding the methylated CG were extracted from the bedMethyle and CG with a score inferior to 5 (Nvalid_cov ) were removed. CG with only one call as modified (Nmod=1) were considered as background and thus considered as unmodified for downstream

analyses. Bw files were generated from bedgraph of the percentage of CGmethylated/CGtotal (%CGmod/CGtotal) and percentage of CG (%CG) using ucsc bedGraphToBigWig (ucsc-bedgraphtobigwig:445--h954228d_0) and visualised with IGV (2.16.2).
The association between the relative level of gene expression and the status of CpG methylation was assessed by isolating the translation start sites (TSS) of protein coding genes from the gene annotation analyses described above. Overlapping genes were excluded from the list of the analysed TSS to limit the generation of artefacts in the downstream analyses. Metaplots and heatmaps of the -300/+200bp surrounding the TSS were performed using the matrix generated with Deeptools (v3.5.6) computeMatrix using the options reference-point --beforeRegionStartLength 300 --afterRegionStartLength 200 -bs 20 --missingDataAsZero --transcript_id_designator "transcript_id". Metaplots and heatmaps were plots using plotHeatmap. A k-means clustering based on the %CGmod/CGtotal was performed using the --kmeans and --clusterUsingSamples options and the regions were sorted based on descending values of %CGmod/CGtotal using --sortRegions descend and --sortUsingSamples options.
Violin plots of the -300/+200bp surrounding the TSS splitted by the k-means algorithm based clusters were generated based on the Metaplot and heatmap datasets using ggplot2 (v3.5.0) and statistical tests were performed using wilcox.test() on R (v4.3.3).

# Data availability

All the raw data produced and used in this study are available in the Bioproject PRJEB75225.
This includes the RNAseq data (ERR12949737), the PacBio Hifi data (ERR14872773), the Hi-C data (ERR14872772), the ONT data (ERR14872784) and the ATAC-seq data (XXXX). The assembly with the functional annotation is available under the accession number XXXX.
The genome resources related to *Xiphinema index* can be viewed at the BIG portal (https://big.plantbios.sophia.inrae.fr).
The full set of analyses are available at https://entrepot.recherche.data.gouv.fr/dataverse/XiphiGeno.

# Acknowledgements

The authors want to thank Christophe Klopp for fruitful discussions and help in Hi-C contact map analysis. Authors are thankful to Carole Belliardo, Georgios D Koutsovoulos and Djampa K L Kozlowski for help in HGT detection, preliminary genome assembly assessment, and basecalling on ONT data, respectively.
All library preparations and sequencings were performed at the GeT-PlaGe core facility (https://doi.org/10.17180/nvxj-5333), INRAE Toulouse. The authors are grateful to the bioinformatics and genomics platform, BIG, Sophia Antipolis (ISC plantBIOs, https://doi.org/10.15454/qyey-ar89) for computing and storage resources and the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) for computing resources.

This work was supported by France Génomique National infrastructure, funded as part of "Investissement d'avenir" program managed by Agence Nationale pour la Recherche (ANR) (contract ANR-10-INBS-09), the French government through the France 2030 investment plan managed by the ANR, as part of the Initiative of Excellence Université Côte d'Azur under reference number ANR-15-IDEX-01 and ANR-13-JSV7-0006 ASEXEVOL for funding the sequencing.

# References

1. Kiontke, K. & Fitch, D. H. A. Nematodes. *Current Biology* **23**, R862–R864 (2013).
2. Bardgett, R. D. & van der Putten, W. H. Belowground biodiversity and ecosystem functioning. *Nature* **515**, 505–511 (2014).
3. Hoogen, J. van den *et al.* Soil nematode abundance and functional group composition at a global scale. *Nature* **572**, 194–198 (2019).
4. van Megen, H. *et al.* A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. *Nematology* **11**, 927–950 (2009).
5. Ahmed, M. *et al.* Phylogenomic Analysis of the Phylum Nematoda: Conflicts and Congruences With Morphology, 18S rRNA, and Mitogenomes. *Frontiers in Ecology and Evolution* **9**, (2022).
6. Qing, X. *et al.* Phylogenomic Insights into the Evolution and Origin of Nematoda. *Systematic Biology* syae073 (2024) doi:10.1093/sysbio/syae073.
7. Holterman, M. *et al.* Disparate gain and loss of parasitic abilities among nematode lineages. *PLOS ONE* **12**, e0185445 (2017).
8. Bird, D. M., Jones, J. T., Opperman, C. H., Kikuchi, T. & Danchin, E. G. J. Signatures of adaptation to plant parasitism in nematode genomes. *Parasitology* **142**, S71–S84 (2015).
9. Danchin, E. G. *et al.* Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 17651–6 (2010).

10. Haegeman, A., Jones, J. T. & Danchin, E. G. Horizontal gene transfer in nematodes: a catalyst for plant parasitism? *Mol Plant Microbe Interact* **24**, 879–87 (2011).
11. Lai, C.-K. *et al.* The Aphelenchoides genomes reveal substantial horizontal gene transfers in the last common ancestor of free-living and major plant-parasitic nematodes. *Molecular Ecology Resources* **23**, 905–919 (2023).
12. Danchin, E. G. J. *et al.* The Transcriptomes of Xiphinema index and Longidorus elongatus Suggest Independent Acquisition of Some Plant Parasitism Genes by Horizontal Gene Transfer in Early-Branching Nematodes. *Genes* **8**, 287 (2017).
13. Qing, X. *et al.* Phylogenomic Insights into the Evolution and Origin of Nematoda. *Systematic Biology* syae073 (2024) doi:10.1093/sysbio/syae073.
14. Esmenjaud, D. Les nématodes de la vigne. in *Ravageurs de la vigne.* 21–45 (Editions Féret, Bordeaux, 2008).
15. Hewitt, W., Raski, D. & Goheen, A. Nematode Vector of Soil-Borne Fanleaf Virus of Grapevines. *Phytopathology* **48**, 586–595 (1958).
16. Andret-Link, P. *et al.* Grapevine fanleaf virus: still a major threat to the grapevine industry. *Journal of Plant Pathology* **86**, 183–195 (2004).
17. Demangeat, G. *et al.* Survival of *Xiphinema index* in Vineyard Soil and Retention of *Grapevine fanleaf virus* Over Extended Time in the Absence of Host Plants. *Phytopathology®* **95**, 1151–1156 (2005).
18. Nguyen, V. C. *et al.* Phylogeography of the soil-borne vector nematode Xiphinema index highly suggests Eastern origin and dissemination with domesticated grapevine. *Sci Rep* **9**, 7313 (2019).
19. Jones, J. T. *et al.* Top 10 plant-parasitic nematodes in molecular plant pathology. *Molecular Plant Pathology* **14**, 946–961 (2013).
20. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**, 484–492 (2012).
21. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: In the right place at the right time. *Science* **361**, 1336–1340 (2018).

22. Brethouwer, T., de Mendoza, A. & Bogdanovic, O. Non-CG DNA methylation in animal genomes. *Nat Genet* **57**, 2395–2407 (2025).

23. Engelhardt, J., Scheer, O., Stadler, P. F. & Prohaska, S. J. Evolution of DNA Methylation Across Ecdysozoa. *J Mol Evol* **90**, 56–72 (2022).

24. Gao, F. *et al.* Differential DNA methylation in discrete developmental stages of the parasitic nematode Trichinella spiralis. *Genome Biology* **13**, R100 (2012).

25. Greer, E. L. *et al.* DNA Methylation on N6-Adenine in *C. elegans*. *Cell* **161**, 868–878 (2015).

26. Liu, Z., Li, Y. & Zhang, X. DNA methylation on C5-Cytosine and N6-Adenine in the Bursaphelenchus xylophilus genome. *BMC Genomics* **24**, 671 (2023).

27. Rošić, S. *et al.* Evolutionary analysis indicates that DNA alkylation damage is a byproduct of cytosine DNA methyltransferase activity. *Nat Genet* **50**, 452–459 (2018).

28. Clapier, C. R., Iwasa, J., Cairns, B. R. & Peterson, C. L. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nat Rev Mol Cell Biol* **18**, 407–422 (2017).

29. Truch, J. *et al.* The chromatin remodeller ATRX facilitates diverse nuclear processes, in a stochastic manner, in both heterochromatin and euchromatin. *Nat Commun* **13**, 3485 (2022).

30. Garrick, D. *et al.* A conserved truncated isoform of the ATR-X syndrome protein lacking the SWI/SNF-homology domain. *Gene* **326**, 23–34 (2004).

31. Argentaro, A. *et al.* Structural consequences of disease-causing mutations in the ATRX-DNMT3-DNMT3L (ADD) domain of the chromatin-associated protein ATRX. *Proceedings of the National Academy of Sciences* **104**, 11939–11944 (2007).

32. Long, H. K. *et al.* Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *eLife* **2**, e00348 (2013).

33. Dalmasso, A. Cytogenetics and Reproduction in Xiphinema and Longidorus. in *Nematode Vectors of Plant Viruses* 139–151 (Springer, Boston, MA, 1975). doi:10.1007/978-1-4684-0841-6_9.

34. Palomares-Rius, J. E., Cantalapiedra-Navarrete, C., Archidona-Yuste, A., Blok, V. C. & Castillo, P. Mitochondrial genome diversity in dagger and needle nematodes (Nematoda: Longidoridae). *Sci Rep* **7**, 41813 (2017).

35. Mota, A. P. Z. *et al.* Unzipped genome assemblies of polyploid root-knot nematodes reveal unusual and clade-specific telomeric repeats. *Nat Commun* **15**, 773 (2024).

36. Yoshida, Y. *et al.* Comparative genomics of the tardigrades Hypsibius dujardini and Ramazzottius varieornatus. *PLOS Biology* **15**, e2002266 (2017).

37. Mitreva, M. *et al.* The draft genome of the parasitic nematode Trichinella spiralis. *Nat Genet* **43**, 228–235 (2011).

38. Korhonen, P. K. *et al.* Phylogenomic and biogeographic reconstruction of the Trichinella complex. *Nat Commun* **7**, 1–8 (2016).

39. Foth, B. J. *et al.* Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction. *Nat Genet* **46**, 693–700 (2014).

40. Lee, Y.-C. *et al.* Single-worm long-read sequencing reveals genome diversity in free-living nematodes. *Nucleic Acids Research* **51**, 8035–8047 (2023).

41. Colinet, D. *et al.* Functional Carbohydrate-Active Enzymes Acquired by Horizontal Gene Transfer from Plants in the Whitefly Bemisia tabaci. *Genome Biology and Evolution* **17**, evaf012 (2025).

42. Ko, I., Kranse, O. P., Senatori, B. & Eves-van den Akker, S. A Critical Appraisal of DNA Transfer from Plants to Parasitic Cyst Nematodes. *Molecular Biology and Evolution* **41**, msae030 (2024).

43. Aguilera, P. & López-Contreras, A. J. ATRX, a guardian of chromatin. *Trends in Genetics* **39**, 505–519 (2023).

44. Villard, L., Fontès, M. & Ewbank, J. J. Characterization of xnp-1, a Caenorhabditis elegans gene similar to the human XNP/ATR-X gene. *Gene* **236**, 13–19 (1999).

45. Dehingia, B., Milewska, M., Janowski, M. & Pękowska, A. CTCF shapes chromatin structure and gene expression in health and disease. *EMBO Rep* **23**, EMBR202255146 (2022).

46. Sawh, A. N. & Mango, S. E. Chromosome organization in 4D: insights from *C. elegans* development. *Current Opinion in Genetics & Development* **75**, 101939 (2022).

47. Halliwell, D. O., Honig, F., Bagby, S., Roy, S. & Murrell, A. Double and single stranded detection of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore sequencing. *Commun Biol* **8**, 243 (2025).

48. Blattler, A. & Farnham, P. J. Cross-talk between Site-specific Transcription Factors and DNA Methylation States*. *Journal of Biological Chemistry* **288**, 34287–34294 (2013).

49. Gibbons, R. J. *et al.* Mutations in ATRX, encoding a SWI/SNF-like protein, cause diverse changes in the pattern of DNA methylation. *Nat Genet* **24**, 368–371 (2000).

50. Voon, H. P. J. *et al.* ATRX Plays a Key Role in Maintaining Silencing at Interstitial Heterochromatic Loci and Imprinted Genes. *Cell Reports* **11**, 405–418 (2015).

51. Hinchie, A. M. *et al.* A persistent variant telomere sequence in a human pedigree. *Nat Commun* **15**, 4681 (2024).

52. Clynes, D. *et al.* Suppression of the alternative lengthening of telomere pathway by the chromatin remodelling factor ATRX. *Nat Commun* **6**, 7538 (2015).

53. Babikir, H. *et al.* ATRX regulates glial identity and the tumor microenvironment in IDH-mutant glioma. *Genome Biol* **22**, 311 (2021).

54. Kernohan, K. D., Vernimmen, D., Gloor, G. B. & Bérubé, N. G. Analysis of neonatal brain lacking ATRX or MeCP2 reveals changes in nucleosome density, CTCF binding and chromatin looping. *Nucleic Acids Res* **42**, 8356–8368 (2014).

55. Maurano, M. T. *et al.* Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Reports* **12**, 1184–1195 (2015).

56. Roseman, S. A. *et al.* DNA methylation insulates genic regions from CTCF loops near nuclear speckles. *eLife* **13**, RP102930 (2025).

57. Eves-van den Akker, S. *et al.* The genome of the yellow potato cyst nematode, Globodera rostochiensis, reveals insights into the basis of parasitism and virulence. *Genome Biology* **17**, 124 (2016).

58. Rancurel, C., Legrand, L. & Danchin, E. G. J. Alienness: Rapid Detection of Candidate Horizontal Gene Transfers across the Tree of Life. *Genes* **8**, 248 (2017).

59. Liu, J. *et al.* The nematode effector calreticulin competes with the high mobility group protein OsHMGB1 for binding to the rice calmodulin-like protein OsCML31 to enhance rice susceptibility to Meloidogyne graminicola. *Plant, Cell & Environment* **47**, 1732–1746 (2024).

60. Bredow, M. & Monaghan, J. Cross-kingdom regulation of calcium- and/or calmodulin-dependent protein kinases by phospho-switches that relieve autoinhibition. *Current Opinion in Plant Biology* **68**, 102251 (2022).

61. Hu, Y. & Hewezi, T. Nematode-secreted peptides and host factor mimicry. *J Exp Bot* **69**, 2866–2868 (2018).

62. Yimer, H. Z. *et al.* Root-knot nematodes produce functional mimics of tyrosine-sulfated plant peptides. *Proceedings of the National Academy of Sciences* **120**, e2304612120 (2023).

63. Nguyen, V. C. *et al.* Evidence of Sexual Reproduction Events in the Dagger Nematode Xiphinema index in Grapevine Resistance Experiments Under Controlled Conditions. *Plant Disease* **105**, 2664–2669 (2021).

64. Blanc-Mathieu, R. *et al.* Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes. *PLOS Genetics* **13**, e1006777 (2017).

65. Dai, D. *et al.* Unzipped chromosome-level genomes reveal allopolyploid nematode origin pattern as unreduced gamete hybridization. *Nat Commun* **14**, 7156 (2023).

66. Jaron, K. S. *et al.* Genomic Features of Parthenogenetic Animals. *Journal of Heredity* **112**, 19–33 (2021).

67. Blanc, C. *et al.* Cosegregation of recombinant chromatids maintains genome-wide heterozygosity in an asexual nematode. *Science Advances* **9**, eadi2804 (2023).

68. Matsuura, K., Fujimoto, M. & Goka, K. Sexual and asexual colony foundation and the mechanism of facultative parthenogenesis in the termite Reticulitermes speratus (Isoptera, Rhinotermitidae). *Insect. Soc.* **51**, 325–332 (2004).

69. Butenko, A., Lukeš, J., Speijer, D. & Wideman, J. G. Mitochondrial genomes revisited: why do different lineages retain different genes? *BMC Biology* **22**, 15 (2024).

70. Puertas, M. J. & González-Sánchez, M. Insertions of mitochondrial DNA into the nucleus—effects and role in cell evolution. *Genome* **63**, 365–374 (2020).

71. Song, H., Moulton, M. J. & Whiting, M. F. Rampant Nuclear Insertion of mtDNA across Diverse Lineages within Orthoptera (Insecta). *PLOS ONE* **9**, e110508 (2014).

72. Villate, L., Esmenjaud, D., Coedel, S. & Plantard, O. Development of nine polymorphic microsatellite markers for the phytoparasitic nematode Xiphinema index, the vector of the grapevine fanleaf virus. *Molecular Ecology Resources* **9**, 229–232 (2009).

73. Van Ghelder, C., Reid, A., Kenyon, D. & Esmenjaud, D. Detection of Nepovirus Vector and Nonvector Xiphinema Species in Grapevine. in *Plant Pathology: Techniques and Protocols* (ed. Lacomme, C.) 149–159 (Springer, New York, NY, 2015). doi:10.1007/978-1-4939-2620-6_12.

74. Hooper, D. & Southey, J. *Laboratory Methods for Work with Plant and Soil Nematodes*. 30 (Ministry of Agriculture, Fisheries and Food, 1986).

75. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).

76. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175 (2021).

77. Brown, M., González De la Rosa, P. M. & Mark, B. A Telomere Identification Toolkit. Zenodo https://doi.org/10.5281/zenodo.10091385 (2023).

78. Dudchenko, O. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).

79. Xu, M. *et al.* TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* **9**, giaa094 (2020).

80. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

81. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

82. Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M. & Blaxter, M. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front. Genet.* **4**, (2013).

83. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**, 1287 (2017).

84. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

85. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**, 1–10 (2020).

86. Baril, T., Galbraith, J. & Hayward, A. Earl Grey: A Fully Automated User-Friendly Transposable Element Annotation and Analysis Pipeline. *Molecular Biology and Evolution* **41**, msae068 (2024).

87. Sallet, E., Gouzy, J. & Schiex, T. EuGene: An Automated Integrative Gene Finder for Eukaryotes and Prokaryotes. *Methods Mol. Biol.* **1962**, 97–120 (2019).

88. Howe, K. L., Bolt, B. J., Shafie, M., Kersey, P. & Berriman, M. WormBase ParaSite - a comprehensive resource for helminth genomics. *Mol Biochem Parasitol* **215**, 2–10 (2017).

89. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **46**, 2699–2699 (2018).

90. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).

91. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protocols* **8**, 1494–1512 (2013).

92. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

93. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

94. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* https://doi.org/10.1093/molbev/msab199 (2021) doi:10.1093/molbev/msab199.

95. Mitchell, A. *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucl. Acids Res.* **43**, D213–D221 (2015).

96. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238 (2019).

97. Solovyev, V. Statistical Approaches in Eukaryotic Gene Prediction. in *Handbook of Statistical Genetics* (John Wiley & Sons, Ltd, 2004). doi:10.1002/0470022620.bbc06.

98. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021).

99. Koutsovoulos, G. D., Noriot, S. G., Bailly-Bechet, M., Danchin, E. G. J. & Rancurel, C. AvP: A software package for automatic phylogenetic detection of candidate horizontal gene transfers. *PLOS Computational Biology* **18**, e1010686 (2022).

100. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772–80 (2013).

101. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).

102. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).

103. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics (Oxford, England)* **17**, 1246–7 (2001).

104. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

105. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

106. Shen, W., Sipos, B. & Zhao, L. SeqKit2: A Swiss army knife for sequence and alignment processing. *iMeta* **3**, e191 (2024).

107. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587–589 (2017).

108. Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol* **59**, 307–321 (2010).

109. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* **35**, 518–522 (2018).

110. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**, W256–W259 (2019).

111. Shimodaira, H. An Approximately Unbiased Test of Phylogenetic Tree Selection. *Syst Biol* **51**, 492–508 (2002).

112. Drula, E. *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Research* **50**, D571–D577 (2022).