Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional

Unidad Irapuato

"Huellas del estilo de vida patogénico en los patrones de selección del genoma del nemátodo *Steinernema carpocapsae*"

Tesis que presenta

M. C. Mitzi Flores Ponce

Para obtener del grado de

Doctor en Ciencias

En la Especialidad de Biotecnología de Plantas

Director de Tesis:

Dr. Rafael Montiel Duarte

Irapuato, Guanajuato Febrero 2017

Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional

Unidad Irapuato

"Footprints of the pathogenic lifestyle in selection patterns in the genome of the nematode *Steinernema carpocapsae*"

Presented by

M. C. Mitzi Flores Ponce

To obtain the

PhD degree

In the specialty of Plant Biotechnology

Thesis advisor:

Dr. Rafael Montiel Duarte

Irapuato, Guanajuato February 2017

Resumen

El nematodo entomopatógeno *Steinernema carpocapsae* se ha empleado mundialmente como agente de control biológico contra plagas de insectos. Esta característica hace que *S. carpocapsae* sea un modelo interesante para entender las interacciones parásito-hospedero. Se han propuesto la acción de dos modelos sobre esta interacción. En un modelo, conocido como "carrera de armamentista", se fijan alelos nuevos por selección direccional positiva en genes relevantes tanto en el hospedero como en los patógenos, produciendo barridos selectivos recurrentes y alternantes. En el otro modelo, conocido como "guerra de trincheras", persisten fluctuaciones dinámicas en las frecuencias alélicas se sustentadas por la selección balanceadora. En este proyecto se investigaron algunos de estos aspectos estudiando la variación genómica en *S. carpocapsae* y otros nematodos de los clados filogenéticos IV y V.

Se realizaron escaneaos del genoma para detectar selección direccional positiva en datos interespecíficos en búsqueda de firmas de la dinámica de carrera armamentista. Los nematodos de vida libre mostraron una proporción significativamente mayor de genes con sitios bajo selección positiva que los nematodos parásitos. Sin embargo, los genes en estos nematodos se muestran un enriquecimiento de categorías funcionales relacionados con la respuesta inmune.

Para detectar posibles efectos de polimorfismos dinámicos se buscaron firmas de la selección balanceadora en datos genómicos intraespecíficos. Utilizando el estadístico Tajima D, observamos la distribución de estos valores a lo largo del genoma. Los valores de Tajima D en *S. carpocapsae* muestran un sesgo hacia valores positivos, significativamente diferentes de la distribución observada en nematodo de vida libre *Caenorhabditis briggsae*.

Finalmente, se evaluó si las proporciones de genes bajo selección se modifican en genes y proteínas diferencialmente expresadas. Se encontró que la proporción de valores significativos de Tajima D positivos se incrementó en genes que expresados diferencialmente después de una inducción con los tejidos de insectos, en comparación con los genes no expresados diferencialmente y el escaneo global.

Este proyecto proporciona un primer vistazo de los efectos que el estilo de vida podría tener en los patrones de selección a nivel genómico. La carrera armamentista entre hospederos y parásitos parece estar afectando a funciones genéticas específicas, pero no necesariamente aumenta el número de genes seleccionados positivamente. La dinámica de la guerra de trincheras parece estar actuando más generalmente en el genoma, probablemente centrándose en los genes que responden a la interacción, en lugar de dirigirse a funciones genéticas específicas.

Abstract

The entomopathogenic nematode *Steinernema carpocapsae* has been used worldwide as a biocontrol agent for insect pests. This feature has make *S. carpocapsae* an interesting model for understanding parasite-host interactions. Two models have been proposed to act in these interactions. In one model, known as “arms race”, new alleles in relevant genes are fixed in both host and pathogens by directional positive selection, producing recurrent and alternating selective sweeps. In the other model, known as “trench warfare”, persistent dynamic fluctuations in allele frequencies are sustained by balancing selection. In these project, we investigate some of these aspects by studying genomic variation in *S. carpocapsae* and other pathogenic and free-living nematodes from phylogenetic clades IV and V.

Genome-wide massive scans to detect directional positive selection in interspecific data was performed to look for signatures of an arms-race dynamic. Free-living nematodes showed a significantly higher proportion of genes with sites under positive selection than in parasitic nematodes. However, genes in parasites presented enriched Gene Ontology terms related to immune response.

To detect possible effects of dynamic polymorphisms interactions we looked for signatures of balancing selection in intraspecific genomic data. Using the Tajima´s D statistic, we observed distribution of Tajima´s D values across the genome. *S. carpocapsae*’s values were more skewed to positive values and significantly different from the observed distribution in the free-living *Caenorhabditis briggsae*.

Finally, we evaluate if the proportions of genes under selection are modified if in the differentially expressed genes and proteins. Finding that the proportion of significant positive values of Tajima’s D was elevated in genes that were differentially expressed after induction with insect tissues as compared to both non-differentially expressed genes and the global scan.

This project provides a first portrait of the effects that lifestyle might have in shaping the patterns of selection at the genomic level. Arms-race between hosts and pathogens seems to be affecting specific genetic functions but not necessarily increasing the number of positively selected genes. Trench warfare dynamics seems to be acting more generally in the genome, likely focusing on genes responding to the interaction, rather than targeting specific genetic functions.

Acknowledgements

To CONACYT for the provided funding (CVU: 267790; Becario:219899).

To CINVESTAV Irapuato and its staff.

To LANGEBIO – UGA and its staff.

To my committee for the patience and the time dedicated to my project.

To Professor Nelson Simões, the external evaluator and colaborator of the project.

To Rafael Montel for embarking with me in this journey. Thank you very much Rafa for the guidance, the patience and all the time it took to plan, discuss and develop the project.

To Cesare Ovando, Christian Martínez Guerrero, Víctor Villa, Ana Juárez, Miguel Vallebueno and Eduardo González for the help with the development of pipelines and software used in this project.

To Hilda Ramos Aboites and Noé García-Chavez for the help with proteomics.

To Manuel Buendía for the help with RNA extraction.

To all the people that passed by Galerón 7. Especially to Hilda and Christian for dealing with us and have the lab running smothly.

To the actual and former members of the Montiel Lab (Lab 13). We always supported each other even though our projects are so diverse.

To the Moreno lab for the support in the final year.

To my dearest friends at Langebio, Karina, Martha, Pablo, Lili, Alex, Sara, Lalo, Miguel, Noé, Brenda, Ana. Thank you for the support and being there for the good and the not so good times.

To all my friends for their loving words and being there when I needed it: Caty, Brianna, Fer, Mariel, Cynthia, Eugenia, Iván. Lulú, Caro, Marian, Mariela, Ale, the Cranfield Mexican crew (especially Alecita, Tere, and Chris), my lovely flat mates (Marc, Jose, Rafa and ChinMay), and Bruno. Thank you very very much.

To my loving family, the biggest thank you of all.Thankyou For all the love and support during my time in grad school.

And to everyone who believed in me even when I didn't.

Thank you!

Index

Figures

Tables

Introduction

Classical evolutionists have been able to infer aspects of the history of living organisms through morphological and physiological comparisons [Nei and Kumar 2000]. However, to achieve a better understanding of the evolutionary history of organisms, other traits need to be analyzed. Advances in biology, specifically in molecular biology and computational technology have improved the way evolution is studied nowadays. Comparisons between organisms at a sequence level is now possible, with DNA sequences the most commonly used [Nei and Kumar 2000].

Molecular Evolution and the neutral theory

The primary cause of evolution are changes in DNA sequences, caused either by nucleotide substitution, insertions/deletions, recombination, or gene conversion [Nei and Kumar, 2000]. Molecular evolution studies how sequences change over time, using two main approaches where hypotheses are tested to search for patterns in the variation of these sequences [Hamilton, 2009; Cutter, 2010]. One of the approaches focuses on specific genes, on the part of the sequence that is changing, and if the change has a functional consequence. The other approach involves testing hypotheses about the population genetic processes that have acted on sequences [Hamilton, 2009].

Molecular evolution's purpose is to distinguish whether patterns of variation are consistent with genetic drift or with certain forms of natural selection [Hamilton, 2009; Ellegren, 2008]. When contrasting the null and alternative hypotheses, the most widely employed null model is based on the neutral theory [Kimura 1968; King and Jukes, 1969]. The neutral theory of molecular evolution postulates that most of the genetic variation within and between species is the result of random factors [Kimura 1968; Nielsen, 2005; Hamilton, 2009]. This process is known as genetic drift, where allele frequencies change randomly in populations dictating the fate (fixation or loss) of these new mutations [Nielsen, 2005; Hamilton, 2009]. Neutral theory adopts the perspective that most observed mutations have little or no fitness effect and are therefore selectively neutral. It does not deny the existence of adaptive mutations but postulates that it

comprehends a minority of the polymorphisms present within a population or the fixations observed between species [Nielsen, 2005; Hamilton, 2009; Fu and Akey, 2013]. Also, it does not deny the appearance of deleterious mutations that will be eliminated from the population by purifying selection, so neutral theory accepts a preeminent role of negative purifying selection [Nei, 2010].

Natural selection

For many years, it has been debated if the genetic variation observed within and among organisms has had a major contribution from adaptive mutations and natural selection. Natural selection occurs when the differences in fitness among individuals produce differential reproduction rates and survival outcomes, leading to variable genotypes [Nielsen, 2005; Hamilton, 2009].

Natural selection cannot only produce an evolutionary change but can maintain genotypes constant [Ridley, 2003]. We can distinguish between three main ways natural selection can act on continuously distributed characters. Selection can be directional when it alters the form of continuous distributions. This form of selection tends to eliminate variation within populations. It can either increase or decrease variation between species [Ridley, 2003; Nielsen, 2005]. Directional selection can be either negative directional selection or positive directional selection. Negative selection or purifying selection can be defined as any type of selection where new mutations decrease the fitness and are selected against, leading to being eliminated from the population [Nielsen, 2005; Duret, 2008; Nei, 2010]. Whereas, in positive selection, certain changes are favored. This is when new mutations confer a higher fitness, becoming advantageous, and tend to increase in frequency over time until they reach fixation [Ridley, 2003; Nielsen, 2005; Duret, 2008].

The second form of selection is balancing or stabilizing selection. This type of selection keeps the population constant through time, and the intermediate allele frequencies have higher fitness than the extremes, therefore variability increases within a population [Ridley, 2003; Nielsen, 2005].

Finally, there is disruptive selection, which occurs when both extremes are favored simultaneously relative to the intermediate types. This type of selection may reduce genetic variability if one of the extreme alleles is fixated [Ridley, 2003; Nielsen, 2005].

It has been proposed that a large proportion of genetic variation is subjected to natural selection and has an impact on organism fitness [Nielsen, 2005]. Therefore, identifying regions that have been shaped by natural selection has become relevant for molecular evolution.

A widely used method to study molecular evolution is a neutrality test. This statistical method explores if the observed genetic diversity is compatible with the neutral evolutionary process, assuming that all mutations are either neutral or strongly deleterious [Nei, 2005; Nielsen, 2005]. There are various statistical methods for testing the neutral theory and aimed to identify regions under selection and distinguish molecular variation that is neutral from variation that is subject to selection, particularly positive selection [Nielsen, 2005]. They could consider either the existence in a population of two or more alleles at one locus, measured by nucleotide diversity (π) for each locus, or divergence estimated by comparing the DNA sequences for both loci between different species, employing a nucleotide substitution model [Nei, 2005; Hamilton, 2009].

Among the approaches to detect positive selection, we identify three major methods testing either divergence or polymorphisms. First, tests looking for divergence among species. These tests evaluate the difference between nonsynonymous mutations per nonsynonymous site (d_N) to the number of synonymous mutations per nonsynonymous site (d_S).The *dN/dS* ratio (or ω) can help identify directional selection [Nei, 2005; Nielsen, 2005].

The second approach is based on polymorphisms within species. These tests examine the pattern of nucleotide frequency distribution and aim to detect directional positive or

negative selection and balancing selection by testing the deviation of the distribution from the neutral expectation. The most commonly used test from this category is Tajima's *D* statistic [Tajima, 1989; Nei, 2005]. The test requires DNA polymorphism data sampled from a single species. Tajima's D compares the average number of nucleotide differences, nucleotide diversity, between pairs of sequences to the total number of segregating sites [Tajima, 1989]. The number of segregating sites is one way to measure polymorphisms in DNA sequences, where a segregating site is any site that maintains two or more nucleotides within a population [Hamilton, 2009]. Under the neutral model, the test expects the difference between the two measures to be zero. Whereas, $D > 0$ suggests balancing selection and $D < 0$ sugest directional selection [Tajima, 1989; Nielsen, 2005; Hamilton, 2009; Nei, 2010].

The third approach is based on both polymorphisms within species and divergence between species. Test falling in this category combine the two approaches described before. The more representative tests are the Hudson–Kreitman–Aguadé (HKA) [Hudson *et al.,* 1987] and McDonald–Kreitman (MK) tests [McDonald and Kreitman, 1991]. The Hudson–Kreitman–Aguadé test considers two or more loci and examines the consistency of variation in both polymorphism within-species and divergence between-species. HKA test requires DNA sequence data from two loci. One locus selectively neutral to serve as a reference and the other locus is the focus of the test. DNA sequence data has its particularities. The sequence of the two loci must be obtained from two species to estimate divergence. Then the test requires sequences from multiple individuals within one of the species to estimate the levels of polymorphism for both loci [Nei, 2005; Nielsen, 2005; Hamilton, 2009]. The McDonald–Kreitman (MK) test, examines whether the ratio of synonymous to nonsynonymous changes within populations is the same as that between populations. MK test requires DNA sequence data from a single coding gene from multiple individuals of a central species to estimate polymorphism, and a DNA sequence at the same locus from another species to estimate divergence [Nei, 2005; Nielsen, 2005; Hamilton, 2009]. Although there are several other tests, they are not used as often as the ones mentioned above.

Studying how natural selection has had an impact on the pattern of variation in species has become interesting, especially with the increasing amount of evidence of positive selection shaping variation within and between species. Some well know examples include the variable region of the immunoglobulin heavy chain in mammals, important for antigen specificity [Tanaka and Nei, 1989]; the RH blood group genes in primates and rodents [Kitano *et al*., 1998; Kitano and Saitou, 1999]; and the Cytochrome c oxidase subunit IV (COX4) gene in primates [Wu *et al*., 1997]; Natural selection has been proposed to be the major force in evolution [Nei, 2005], generating adaptation when the environment changes, additionally, it can explain both the evolutionary change and the absence of it [Ridley, 2003]. Also, studing the adaptation ocurring on interacting species becomes interesting, considering the possibility that their relationship can contribute to their variation patterns, where the interaction exerts reciprocal natural selection on each other, leading to coevolution.

Coevolution

Coevolution is the process of reciprocal adaptive change in the genetic composition of two or more species, where one species changes in response to the change in the interacting species [Woolhouse *et al*. 2002; Jackson, 2008]. This coevolutionary co-adaptive dynamic can occur between any interacting populations, with one of the most closely studied being the host-pathogen interaction [Woolhouse *et al*. 2002].

Hosts and pathogens compete or interact in such a way that an equilibrium is never reached [Nielsen, 2005], with hosts evolving under selective pressure to avoid pathogen infection and pathogens with the pressure to evade host defences [Aguileta *et al.*, 2009]. Coevolution between hosts and their parasites occurs under a wide range of conditions with many outcomes, and it is expected that coevolution follows a range of possible dynamics. The two extreme coevolutionary dynamics involve selective sweeps and dynamic polymorphisms [Woolhouse *et al*., 2002; Tellier *et al*., 2014]. Selective sweeps occur when new alleles appear, by mutation or migration, and eventually become fixed in the population by directional positive selection. The recurrent fixation of alleles is called an “arms race” [Woolhouse *et al*., 2002; Tellier *et al*., 2014]. On the other hand,

dynamic polymorphisms involve continuous cycling in allele frequencies and are evolving under balancing selection. Such fluctuations caused by selection are inherently persistent, although fixation can occur as a result of genetic drift. This model is known as “trench warfare” or “Red Queen” and has been predicted to exhibit faster coevolutionary cycles than arms race [Woolhouse *et al*., 2002; Aguileta *et al*., 2009; Tellier *et al*., 2014].

Well-known examples of a coevolutionary “arms race” dynamic are genes involved in immunity and defense, which are expected to evolve under positive selection [Nielsen, 2005; Aguileta *et al*., 2009]. Such is the case of the *scr*74 gene family in *Phytophthora infestans* that encodes phytotoxins, small secreted peptides that induce cellular death in their host [Aguileta et al., 2009]. Another example is the primate gene apolipoprotein B–editing catalytic polypeptide 3G (APOBEC3G), which encodes for a cytoplasmic protein that catalyze the deamination of cytosine to uracil DNA and RNA, and provides differential susceptibility to HIV infection and simian immunodeficiency virus (SIV) [Sawyer et *al.,* 2004; Nielsen et *al.,* 2005]. Plant R genes are other examples of genes evolving under positive selection in response to host-pathogen interactions. In tomato, R-gene loci Cf-4/9 and Cf-2/5 provide resistance to the fungal parasite *Cladosporium fulvum*, [Stahl et *al*., 2000; Nielsen*,* 2005]. Other examples are the transmembrane protein PorB of *Meningococcu* and the capsid genes of the Canine parvovirus (CPV). Both these surface proteins are probably involved in the evolutionary arms race between hosts and pathogens [Aguileta *et al*., 2009].

However, it is not clear to what extent the coevolutionary interaction might alter the evolutionary patterns at the genome level, or to what extent it might affect levels of intraspecific diversity. One expectation is that the total number of genes with specific signatures of selection, either from positive or balancing selection, would increase in comparison to genomes of non-pathogenic organisms. An example of this expectation is a study in 2013 where plant genomes revealed that parasitic lineages have a faster rate of molecular evolution than their non-parasitic relatives [Bromham *et al*., 2013]. While this result is not conclusive and it does not detect positive selection, it suggests

that a cause of the raised non-synonymous substitution rate could be positive selection, although it does not discard other demographic effects [Bromham *et al*., 2013].

Another expectation would be that the number of genes with signals of selection will increase in genes participating in the interaction as compared with genes that do not. It might be difficult to find all the genes involved in the host-pathogen interaction, but a first approximation can be obtained by inducing the pathogen with host tissues and identifying the differentially expressed genes [Dobon *et al*., 2016].
If the host-pathogen interaction is increasing the number of genes evolving under positive selection, responsible for an increased number of selective sweeps, then a reduction in diversity levels is expected in pathogen when compared with non-pathogen genomes. On the contrary, if the effects of the interaction are more related with balancing selection, then the above mentioned reduction in diversity will not be found.
The most likely scenario is that part of the genes might be under positive selection and part of them under balancing selection, canceling the effect of each other in diversity. In any case, evolutionary genetic interactions might represent an additional determinant of genetic diversity.

Parasitism in nematodes

Nematodes are the most abundant type of animals on earth in terms of the number of individuals, being an ancient and diverse group [Platt, 1994]. Their diversity and abundance are the results of their extraordinary ability to adapt, their small size, resistant cuticle, and simple body plan [Coghlan, 2005].

Many types of associations exist between nematodes and insects, ranging from phoresis to parasitism and pathogenesis [Smart 1995; Giblin-Davis *et al*., 2013; Blaxter and Koutsovoulos, 2015]. All these symbiotic interactions involve different mechanism. Phoresis involves the transport of one organism by another, with no physiological or biochemical dependence between the host and symbiont [Clausen 1976]. A parasitic symbiotic relation involves a host and a parasite (or pathogen if the disease occurs in the host) [Davis *et al*., 2000]. In this relationship, a parasite can be defined as an

organism living in or on another living organism, obtaining from it part or all of its organic nutriment, and commonly exhibiting some degree of adaptive structural modification [Davis *et al*., 2000]. Consequently, a parasite can spend a significant portion of its life in or on the living tissue of a host organism and cause harm to the host without immediately killing it [Yarwood 1956; Davis *et al*. 2000].

In nematodes, parasitism has arisen independently multiple times from free-living species to parasite plants and animals [Dorris *et al*., 1999; Blaxter and Koutsovoulos, 2015]. Invertebrate parasitism, specifically parasitism of insects, has evolved independently at least twice [Blaxter *et al.*, 1998; Blaxter and Koutsovoulos, 2015]. Recent phylogenetic associations between these invertebrate parasites and vertebrate parasites suggest that entomopathogenic nematodes were the ancestor from which vertebrate-parasitic nematodes evolved [Blaxter and Koutsovoulos, 2015].

It is believed that the change to parasitism probably required the adaptation of genes present in their free-living ancestors [Blaxter 2003; Coghlan 2005], through the action of positive selection. These genes may be manifested as morphological structures that provide access to parasitism of a specific host or they may be involved in the interaction with its host. They could also provide the ability to survive the immunological attack, or to develop strategies to overcome the host defenses, such as evasion, tolerance and suppression [Boemare *et al*., 1996; Davis *et al*., 2004; Coghlan, 2005].

Entomopathogenic nematodes

Entomopathogenic nematodes (EPN) are lethal pathogens that normally kill the insect host and develop in the resulting cadaver. These obligate pathogens contribute to the regulation of natural populations of insects and can infect a wide range of insects and have a common morphological feature, the infective juvenile (IJ) or dauer juvenile [Smart 1995; Castillo *et al.,* 2011]. The IJ is a non-feeding stage required for the successful infection of insect host [Castillo *et al.,* 2011].

There are many genera of nematodes that parasite insects. However, the EPN research

is largely concentrated and mainly refers to only two families of rhabditid nematodes: Steinernematidae and Heterorhabditidae [Poinar, 1979; Burnell and Stock, 2000]. Species belonging to these two families provide effective biological control of a variety of economically important insect pests, having as their natural host the soil-dwelling stages of lepidopteran, dipteran and coleopteran pests of commercial crops [Smart, 1995].

EPNs possess several attractive biotechnological qualities for agriculture besides their broad host range. Such qualities include high pathogenicity, a durable infective stage, host-seeking ability, safety to non-target organisms, (including vertebrates, plants and other insects; bacterial association under natural conditions, ease to mass production, compatibility with many chemical pesticides and are amenable to genetic selection [Kaya and Gaugler, 1993; Griffin *et al*., 2005; Li *et al*., 2009]. These qualities have given them an important role regulating the productivity of wild soil-dwelling populations, which have a negative impact on human agricultural production and offer an alternative to chemical insecticides. Currently, EPNs are being commercialized worldwide for biological control of several insect pests [Blaxter, 2003].

Additionally, EPNs have a specific association with two genus of enteric bacterium, which inhabit the intestine of the nematode. This symbiotic relationship (*Photorhabdus* spp. for *Heterorhabditids* and *Xenorhabdus* for Steinernematids) make them unique among rhabditids [Boemare, 2002; Hao *et al*., 2008; Mbata and Shapiro-Ilan, 2010]. They also share a similar life cycle through convergent evolution and are closely related to vertebrate parasitic nematodes such as Strongylida and Rhadbditida (Strongyloididae) [Li *et al*., 2009].

Steinernema carpocapsae

Steinernema carpocapsae is one of the most well-known species of EPNs. It associates symbiotically with the gram-negative enterobacterium *Xenorhabdus nematophila* to form a pathogenic complex capable of parasitizing and killing a broad range of insects and is commercially produced as a biological control agent against insect pests [Kaya and

Gaugler, 1993; Ehlers 2001; Balasubramanian 2009].

S. carpocapsae has a very simple life cycle compared to other nematodes. It consists of adults, embryos and four juvenile stages [Smart, 1995; Shapiro-Ilan and Gaugler, 2002]. The third juvenile stage can shift to a free-living non-feeding stage, the IJ, in response to environmental nutritional cues [Kaya and Gaugler, 1993; Gaugler, 2002] IJs retain the previous stage cuticle, which assists them to survive in the soil outside the host [Smart 1995; Castillo *et al.,* 2011]

The lifecycle resumes after the *S. carpocapsae* IJs actively search for the insect host. Once found, they enter through natural openings, reach the hemocoel, and release their symbiotic bacteria causing septicemia and subsequent death of the host within 48 hours [Smart, 1995; Burnell and Stock, 2000]. *X. nematophila* provide nutrients for themselves and the nematode by secreting a variety of extracellular enzymes which degrade host tissue. Then the nematodes feed on bacteria in the cadaver, reproduce and complete their life cycle, generally completing two to three generations in the same cadaver [Burnell and Stock, 2000]. Once the supply of nutrients is depleted, the IJs emerge from the cadaver and can seek new hosts [Gaugler, 2002].

For a long time, *S. carpocapsae* and the rest of the EPNs were only seen as bacterial vectors. However, there is increasing evidence showing that the nematode has a more active role during pathogenesis [Kaya and Gaugler, 1993; Smart, 1995; Simões *et al*., 2000; Hao *et al*., 2008; Castillo *et al*., 2011]. In fact, reports in the early 80s showed that the mutualistic interaction of *S. carpocapsae-X. nematophila*'s is not obligated, and that axenic *S. carpocapsae* can infect, interact with and parasite insects [Gotz *et a*l., 1981; Burman, 1982]. It was also proposed that possible virulent factors could be the proteases secreted during the initial development of the nematode inside its host [Simões and Rosa, 1996]. In 2000, axenic *S. carpocapsae* was proved to be able to destroy antibacterial factors and developed toxemia conditions in parasitized insects [Simões *et al*., 2000]. Recent genomic studies in *S. carpocapsae* [Dillman *et al*., 2015; Rougon-Cardoso *et al*., 2016] have shown specific evolutionary and functional signatures in its genome that can be related to parasitism. These involve a set of expanded gene families

likely involved in parasitism, orthologous genes shared with other parasitic nematodes not present in free-living species, ncRNA families reported to be enriched in parasites, and the expression of proteins putatively associated with parasitism and pathogenesis [Dillman *et al*., 2015; Rougon-Cardoso *et al*., 2016]. These signatures are most likely the result of evolutionary interaction with the hosts and suggest an active role during the pathogenic process.

Given the parasitic nature of *S. carpocapsae*, it seems interesting to study how selection pressures related to its lifestyle have affected its genome. We looked for ancient patterns of selection through interspecific data, including nematodes with different lifestyles, and evaluated more recent signatures of selection through intraspecific data. Also, we evaluated if patterns of selection were modified in differentially expressed genes after induction with insect tissues.

Hypothesis

The genome of *Steinernema carpocapsae*, as a parasitic nematode, will have differential selection patterns in comparison with free-living nematodes.

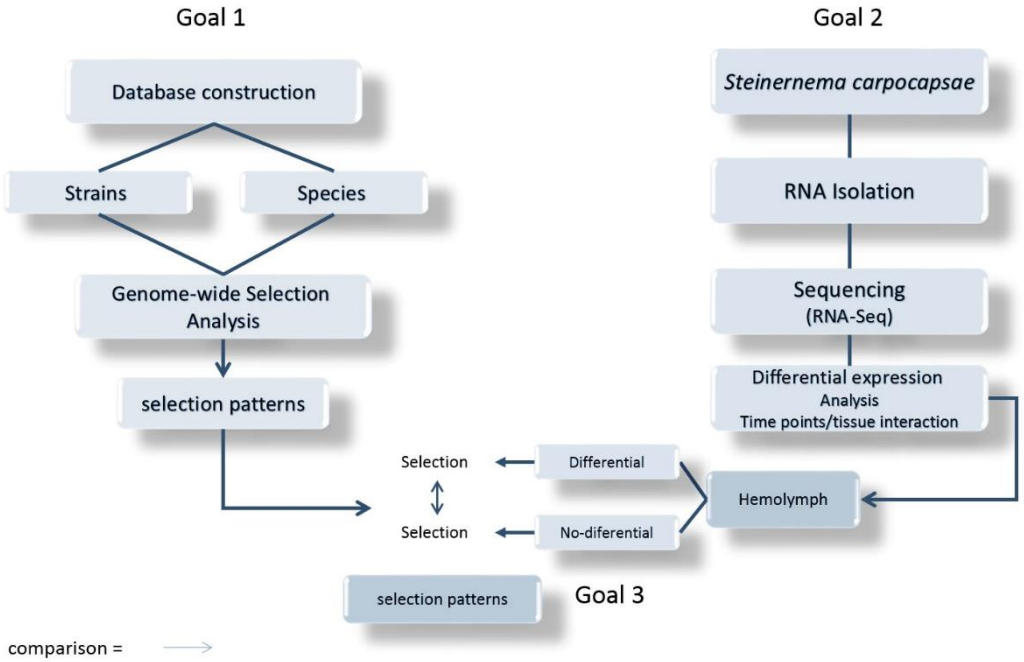
Goals

Determine how the nematode *S. carpocapsae* genome has responded to a parasitic life style in evolutionary terms

Specific goals

1. Determine the proportion of the *S. carpocapsae* genome that is under selection.
2. Look for differentially expressed genes due to the interaction with insect tissue.
3. Determine if the proportions of genes under selection are modified if in the differentially expressed genes.

Experimental Strategy



Materials and Methods

Interspecific analysis

Species and sequences

A group of eight parasitic and free-living nematode species were chosen. Nematodes belonged to clades IV and V acording to Blaxter *et al*. [Blaxter *et al.,* 1998]. Parallel analysis of each clade allowed an independet evaluation of the effects in the genome of different life styles. Nematodes from clade IV included the free-living nematode *Panagrellus redivivus*, the vertebrate parasite *Strongyloides ratti*, the entomopathogenic *Steinernema carpocapsae*, and the mycophagous *Bursaphelenchus xylophilus*. Nematodes from clade V included the free-living nematodes *Caenorhabditis briggsae* and *Pristionchus pacificus*, the vertebrate parasite *Haemonchus contortus*, and the entomopathogenic nematode *Heterorhabditis bacteriophora*. Although *B. xylophilus* is a fungal feeding nematode that colonize dead or dying trees [Jones *et al.,* 2008], rather than a strictly free-living nematode, we incorporated it into our dataset to have a fair number of species and to try to balance the datasets between clades.

All nucleic and amino acid sequences of seven of the nematode species (*C.briggsae, Pristionchus pacificus, B. xylophilus, Heterorhabditis bacteriophora, Haemonchus contortus, Panagrellus redivivus, Strongyloides ratti*) were downloaded from Wormbase (ftp.wormbase.org release WS244 1-Sept-2014) [Harris *et al*., 2010]. Sequences from *S. carpocapsae* strain Breton were generated previously in our group (GenBank Bioproject ID# 39853) [Rougon-Cardoso *et al.,* 2016].

The headers and names of the protein and CDS files were checked to match with the annotation files (.gff). In the cases where the species identifier was not specified (i.e gene_001, protein_001) we renamed them with the first letters of the organism genus and species (i.e. Sr_gene001 and Sr_protein001 for *Strongyloides ratti*). Finally, the rest of the headers were renamed with a shorter name to avoid problems in the rest of the analysis.

Orthologous genes

Nematode species were grouped into three different data sets (Table 1), two sets of four nematodes from each clade, and with all the species. Orthologue genes were identified using amino acid sequences for each set, with OrthoMCL v.2.0.7 [Li *et al.*, 2003] using an inflation index of 1.5. This parameter regulates the group cluster tightness and appears to balance sensitivity and selectivity at a value of 1.5 [Li *et al.*, 2003].

Table 1. Nematode sets for orthologue searches

Set name	Number of species	Species	Clade *
C4N4	4	*Bursaphelenchus xylophilus, Panagrellus redivivus, Steinernema carpocapsae and Strongyloides ratti*	Clade IV
C5N4	4	*Caenorhabditis briggsae, Heterorhabditis bacteriophora, Haemonchus contortus and Pristionchus pacificus*	Clade V

Sets were named after the clade they belong to and the number of species included. *Classification according to Blaxter *et al.,* 1998.

Scans of positive selection

Amino acid and nucleotide sequences from the orthologous genes were aligned with Clustalw2 v.2.1 [Thompson *et al.*, 1994; Larkin *et al.,* 2007]. Then, the nucleotide alignments used in the selection scans were obtained with RevTrans v.1.4 [Wernersson and Pedersen, 2003] based on the complete amino acid alignment to preserve codon homology.

For the phylogenetic reconstruction, the informative blocks were recovered with Gblocks [Castresana, 2000] and the best-fit evolutionary model for each of the aligned protein was obtained with ProtTest [Darriba *et al.,* 2011]. Then sequences were grouped by their evolutionary model and concatenated. We then reconstructed a consensus phylogenetic tree with the aLRT implementation in PhyML v.3.0 [Guindon and Gascuel, 2003].

The trees and nucleotide alignments were used to assess signatures of natural selection with Codeml from the PAML package v.4.6 [Yang, 2007]. Genome-wide scans of positive selection were performed with the help of the program Clarisse, which automatizes the Codeml runs, letting us test several models and hypothesis with Likelihood Ratio Tests (LRTs) in a more simple way. This program was designed and kindly provided by M.Sc. Victor Villa Moreno, from the "Laboratorio de la Diversidad Biomolecular" (Langebio) under the direction of Dr. Mauricio Carrillo Tripp. Clarisse uses two files, the configuration file in which the user indicates the parameters that will change for each run. The second is the template file, which has the parameters that will not change through all the codeml runs (See appendix 1).

Codeml calculates ω (dN/dS), the ratio between the rate of nonsynonymous mutations per nonsynonymous site (dN) and the rate of synonymous mutations per nonsynonymous site (dS) [Yang, 1998; Yang, 2000]. Where ω >1 is indicative of positive selection, ω = 1 corresponds to the neutral expectation and ω < 1 indicates negative or purifying selection. Codeml also estimates several parameters used to calculate ω. These include the sequence divergence (t), and the transition/transversion rate ratio (k).

Two models, branch and branch-site, were used to identify genes and sites under directional positive selection, using LRTs to assess significance. LRTs consist in the comparison of twice the log-likelihood difference between the two models tested with a χ^2 distribution with degrees of freedom (df), equal to the difference in the number of parameters for the models tested [Yang and Bielawski, 2000]. For for most of the tests df=1. The only test with a different value was the M0-M1 test, with df=4. The branch model is useful for detecting positive selection acting on particular lineages [Yang 1998; Yang and Nielsen 1998]. Whereas the branch-site model intends to detect positive selection affecting just a few sites along particular lineages [Yang and Nielsen, 2002; Zhang *et al*., 2005]. LRTs performed for each model are shown in Table 2.

Table 2. Likelihood Ratio Tests (LRTs) performed for each selected branch

Codon substitution Model	**LRT**	**Models**	**Parameters in codeml**
Branch	H_0: Same ω for all branches H_1: Different ω for all branches	Model 0: one-ratio model vs Model 1: free-ratios model.	Model 0: model = 0, fix_omega = 0, omega = 0 Model 1: model = 1, fix_omega = 0, omega = 0
Branch	H_0: Same ω for all branches (background) H_1: A different ω for the selected branch (foreground)	Model 0: one-ratio model vs Model 2: different ratio in the specified branch	Model 0: model = 0, fix_omega = 0, omega = 0 Model 2: model = 2, fix_omega = 0, omega = 0
Branch	H_0: ω = 1 for the foreground branch H_1: ω ≠ 1 for the foreground branch	Model 2: different ratio in specified branch vs Model 2 fix ω: ratio=1 in the specified branch	Model 2: model = 2, fix_omega = 0, omega = 0 Model 2 fix ω: model = 2, fix_omega = 1, omega = 1
Branch-Site	H_0: Same ω for all sites among branches. H_1: Different ω for all sites in the foreground branch.	A: different ratio per site in specified branch vs A1: ω ratio=1 per site in the specified branch	A: model = 2, NSsites = 2, fix_omega = 0 A1: model = 2, NSsites = 2, fix_omega = 1, omega = 1

NSsites = 0 for all the branch models. kappa was estimated for each gene and fixed with fix_kappa = 1 and kappa = estimated value. For all the models these parameters were the same: noisy = 3, verbose = 0,runmode = 0, seqtype = 1, CodonFreq = 2, clock = 0, aaDist = 0, icode = 0, fix_alpha = 1, alpha = 0, Malpha = 0, ncatG = 10, getSE = 0, RateAncestor = 0, Small_Diff = .5e-6, cleandata = 1, method = 1.

The significance of differences in the proportion of genes with sites under selection between parasitic and free-living nematodes was assessed with a χ^2 test on a 2x2 contingency table.

Functional analysis and over-represented functional categories in positively selected genes

Functional annotation and enrichment of functional categories in genes with positively selected sites were performed with Blast2GO [Conesa *et al.* 2005]. Blast2Go was used to look for protein functional prediction based in a homolgy search with BLAST against the NR database. Enrichment consisted in the identification of how many genes were rated to at least one GO category and the performance of the Fisher's exact test comparing the genes with sites under positive section with the rest of the orthologue gene sets.

Intraspecific analysis

Species and sequences

Strains of *S. carpocapsae* and *C. briggsae* [Thomas *et al.* 2015], were used for the population genetics approach. Strains details are shown in Table 3. *S. carpocapsae* strains were cultured in our lab.

Table 3. *S. carpocapsae* and *C.briggsae* strains used for intraspecific analysis.

Species	Strain	Geographic origin	Sequencing strategy	Source	Reference
S. carpocapsae	Breton*	France	454flx, SOLiD	NS	Rougon-Cardoso *et al.* 2016
S. carpocapsae	All	USA	HiSeq 2500	HGB	This study
S. carpocapsae	Az20	Açores, Portugal	HiSeq 2500	NS	This study
S. carpocapsae	Az154	Açores, Portugal	HiSeq 2500	NS	This study
S. carpocapsae	Az157	Açores, Portugal	HiSeq 2500	NS	This study
C. briggsae	AF16*	Gujarat, India	Combined	Wormbase	Stein *et al.* 2003
C. briggsae	JU1348	Kerala, India	HiSeq 2000	NCBI SRR1793004	Thomas *et al*. 2015
C. briggsae	QR25	Quebec, Canada	GA IIx	NCBI SRR1793006	Thomas *et al*. 2015
C. briggsae	VX0034	Hubei, China	HiSeq 2000	NCBI SRR1793007	Thomas *et al*. 2015
C. briggsae	ED3101	Nairobi, Kenya	GA IIx	NCBI SRR1793002	Thomas *et al*. 2015

Strains are natural isolates. Sequencing strategy: 454flx, Roche pyrosequencing; HiSeq, Illumina; Combined, whole-genome shotgun sequencing (WGS) with a high-resolution, sequence-ready physical map; GA, Genome Analyzer. Libraries were Paired End (2x100bp).Strain sources are: NS, Nelson Simões; HGB, Heidi Goodrich-Blair. NCBI, National Center for Biotechnology Information. *Reference genome; *S. carpocapsae* from [Rougon-Cardoso *et al.,* 2016], and *C. briggsae* from wormbase.org (c_briggsae.PRJNA10731.WS253).

Total DNA from the four strains of *S. carpocapsae* was extracted from a pool of nematodes using a phenol/chloroform extraction protocol described in Sambrook *et al.* [Sambrook *et al*., 1989]. DNA yield and integrity was measured with a 2100 Bioanalyzer (Agilent) using an Expert High Sensitivity DNA chip and sequenced with the Illumina HiSeq 2500 platform, at the Cinvestav-Langebio Core Facility.

Sequences from *C. briggsae* were downloaded from the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) and converted to fastq files with SRA Toolkit v.2.3.5-2 (The NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra).

The ML genetic distance between the *S. carpocapsae* strains and *C. briggsae* strains was calculated with Tree-Puzzle v.5.3.rc16 [Schmidt *et al*., 2002], using the best-fit models estimated with jModelTest v.2.1.3 [Darriba *et al.,* 2012].

Mapping and SNP determination

The reads of the *S. carpocapsae* and *C. briggsae* strains were mapped using the Burrows-Wheeler Aligner (BWA) v. 0.7.12 MEM algorithm [Li and Durbin 2009; Li 2013] to the *S. carpocapsae*'s genome sequenced in our group (GenBank Bioproject ID# 39853); and the reference genome assembly PRJNA10731 (release WS253) of *C. briggsae* strain AF16 (http://www.wormbase.org), respectively. Commands and mapping parameters are shown in Appendix 1.

Besides mapping, the alignments were filtered with the awk command. Then, SAM files were processed using samtools v. 1.3.1 [Li *et al*. 2009], by (***view)***, then sorted (***sort***) and indexed (***index***). Finally, potential PCR duplicates reads were removed (***rmdup)***. SNP callings were conducted with a pipeline which included the use of samtools v. 1.3.1

[Li *et al* 2009] ***mpileup***, that generates a position-based output, a consensus/indel calling, and bcftools v. 1.2 [Li 2011] ***call***, to generate a SNP/indel variant calling from the BCF file. Parameters are shown in Appendix 1.

The outputs of the pipleline, Variant Call Format (vcf) files, were used to construct haplotype maps files in the format used by the International HapMap Consortium for each of the nematode species. This consisted of the alignment of the vcf files to the reference genomes used in the mapping section. Regions with repetitive elements or with a lack of coverage were filtered out.

Selection scans and statistical tests

For the selection analysis between species, we used Tajima's D, a pymorphism-based test. This test studies the relationship between the number of segregating sites and the average number of nucleotide differences estimated from pairwise comparisons [Tajima 1989].

In order to perform genome-wide massive selection scans, we used a program called "Massive Tajima" developed in our group. This program was used to estimate the index of nucleotide diversity (π) [Nei, 1990] and theta (θ) [Hamilton, 2009] to calculate the neutrality estimator Tajima's D [Tajima, 1989] in non-overlapping sliding windows. Additionally, this program is capable of generating and analyzing the observed distribution of D values. The analysis was done across the complete genomes in windows of 1,000 bp. Windows without variation (π =0) were marked as invariable and excluded from the distribution. In these windows, a zero is obtained in the denominator and therefore the estimation of Tajima's D is not possible. Also, windows with less than 90% of coverage in the haplotypes maps files were not taken into account for the analysis.

The average and the observed distribution of D values for each species was obtained and used to calculate corrected confidence limits [Schmidt and Pool, 2002]. A determined percentile on each tail of the distribution was used to find the confidence limits to identify windows with significant D values. To achieve a confidence of 95%, we

intended to record the 2.5 percentile on each tail of the observed distribution. Since D values were non-continuous (i.e. were distributed in discrete intervals) we selected the 2 percentile on each tail (otherwise we would have needed to jump to the 3 percentile). By selecting only the 2% of each tail we exclude with 96% confidence the effect of demoghraphic proceses in Tajima's D value, which would be explained only by selection forces in these cases.This is because of demographic proceses are expected to afect the whole genome with a similar strength and only the most extreme values will be free of this effect. However, we are probaly excluding other genes also affected by selection, present in the 2-3% interval. Therefore the 2% most positive values and the 2% most negative values were considered as candidate regions to have been targeted by selection, under the corrected confidence limits of the distribution. Finally, we compared the observed confidence limits to the theoretical confidence limits reported by Tajima using a theoretical data distribution [Tajima, 1989].

Differentially expressed genes after interaction with insect tissue

Organisms, maintenance, and storage

S. carpocapsae strain Breton were obtained from Dr. Nelson Simões from Azores University, Portugal. The culture was routinely maintained through the infection of *Galleria mellonella* (wax moth) larvae at 25°C [Kaya and Stock, 1997]. For mass production, nematodes were grown using a modified protocol in an artificial medium according to Bedding [Bedding, 1981]. Nematodes were stored as IJs at 10°C in sterile tap water at a concentration of approximate 5,000 nematodes/ml.

Infection kinetics

In order to identify a more specific time when *S. carpocapsae's* IJs enter the insect's internal organs, we designed an experiment to estimate the time when IJ reaches the insect's intestines and hemocoel [Adapted from Simões *et al.,* 2000]. We measured and

weighed *G. mellonella* larvae, and set them in groups of three to work with batches. All batches had similar length and weight (approximately 2 cm and 220 mg) (Figure 1).

Figure 1. Measuring *Galleria mellonella* larvae. All the larvae used in the infection kinetics were measured to have a consistent group of individuals.

Larvae were individually placed on a 3.8 cm^2 well, from tissue culture plates, with double filter paper. Then 100 ml of IJs suspension at a concentration of 45 nematodes per ml, were added to each well. The plate was incubated at 23°C protected from light for 3, 4, 5, 8 and 10 h. We did three batches per time point, with a delay among them to have time for the larvae dissection. Just before dissections, we rinsed the larvae to remove nematodes that could be attached to the surface. This was done passing the larvae in miliQ water two times (different containers) then dried with a paper towel [Adapted from Simões *et al.,* 2000]. Larvae were pinned to a wax base, then dissected cutting ventrally carefully to avoid cutting the intestine. Once open, the complete intestine was carefully removed and placed into a petri dish with saline solution. The rest of the body was placed on a different petri dish also with saline solution. We took care not to spill the hemolymph and tissues on the wax base to avoid losing nematodes. Nematodes were then counted under the microscope.

Hemolymph extraction

Hemolymph used in our experiments came from whole *G. mellonella* larvae. Larvae were frozen with liquid nitrogen and homogenized in a blender using a small container. Mashed larvae were transferred to 50 ml plastic tubes containing the same volume of cold Tyrod buffer 1X (NaCl 0.8%, KCl 0.02%,CaCl2 0.02%, MgCl2, 0.02%, NaH2PO4, 0.005%, NaHCO3, 0.1%, and glucose, 0.1%). They were then sonicated for 5 minutes, ad further centrifuged at 4000 rpm for 15 minutes at 4ºC to form three layers. Hemolymph constituted the middle layer that was transferred with a pipette into microtubes in 500 µl aliquots and used immediately or stored at -20ºC [modified from Toubarro *et al*. 2013].

Nematode induction with insect (*G. mellonella*) hemolymph

To simulate the infection process, induction with insect (*G. mellonella*) hemolymph was performed. Approximately 24,000 nematodes (5 ml of IJs suspension) were passed on 50 ml falcon tubes. The IJs were washed three times stirring them for 10 minutes, followed by a 2 minute centrifugation at 2000 rpm at 10°C. The first wash was with a solution of 2% sodium hypochlorite, the subsequent washes were only with sterile tap water or sterile saline solution 0.8%. Once the washing process was done, nematodes were placed under sterile environment to avoid external contamination, and the liquid was removed by filtration. Filtration was performed placing filter paper on a filtration unit connected to a vacuum pump. When the nematodes were slightly dry, the filter paper was placed into a petri dish containing 4.5 ml of Tyroid buffer 1X, 50 µl of nalidixic acid (25µg/µl), and 450 µl of insect hemolymph. Nematodes were then incubated 1 and 2 hours at 23-25ºC with orbital stirring at 50 rpm [Simões *et al.,* 2000; Hao *et al*., 2010]. Ampicillin was changed from the original protocol for nalidixic acid, which gave a better response in killing *X. nematophila* in an antibiogram. Once the incubation time was over, we checked for nematode vitality taking a small drop and looking for nematode motility under the microscope. Nematodes were collected with sterile filter paper in the filtration unit under the laminar flux hood, then quickly washed 3 times with 15 ml of sterile cold (10 ºC) miliQ water. Dry nematodes were then grinded under liquid nitrogen and stored

at -70ºC. We took extreme care to maintain the nematodes frozen at all times. Samples were later used for RNA isolation.

RNA isolation and sequencing

Total RNA was extracted from the induced IJs using the TRIzol (Invitrogen) protocol. Protocol started with a volume of TRIzol reagent was added and incubated for 5 minutes at room temperature. Then 0.2 ml of chloroform was added, followed by vigorously shaking for 15 seconds. After this, the tube was incubated for 2-3 minutes at room temperature, centrifuged at 12,000 × g for 15 minutes at 4°C. Once the centrifugation was over, the aqueous phase was removed and placed into a new tube. Then 0.5 ml of isopropanol 100% were added and incubated 10 minutes at room temperature. Following the incubation, tubes were centrifuged 12,000 × g for 10 minutes at 4°C. After centrifugation, the supernatant was removed, and 1 ml of ethanol 75% was added to wash the RNA pellet. The tube was vortexed briefly and centrifuged at 7500 × g for 5 minutes at 4°C. Supernatant was discarded and the pellet dried for 5–10 minutes. The pellet was later suspended in RNase-free water (20–50 µl). Later both yield and quality were verified by a 2100 Bioanalyzer (Agilent). Paired-end (PE) TruSeq RNA libraries were constructed at the Genomic Services of Langebio. Sequencing was done in two places. At LANGEBIO core facilities with the MiSeq, HiSeq 2500 (illumina) platforms, and at UC Davis Genome Center sequence facilities with the HiSeq 3000 (illumina) platform.

Differential expression analysis

For the analysis of gene expression, we used a group of third-party software, set into a pipeline disposition with bash and R scripts. The pipeline design (Figure 2) was done in collaboration with Dr. Cesare Ovando-Vázquez. All the work was done in the Mazorka computational cluster at Langebio.

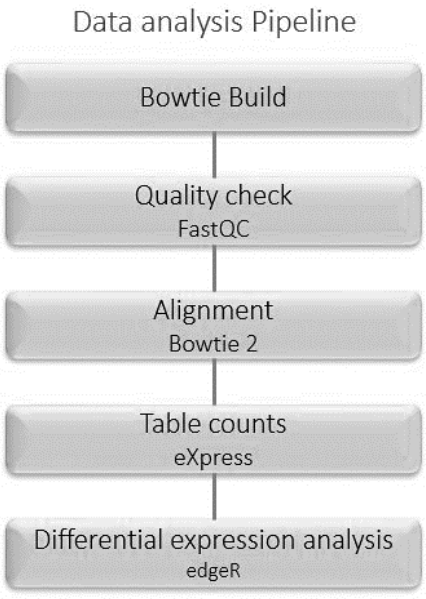


Figure 2. Differential expression analysis pipeline. Description of the main steps in the gene expression analysis pipeline.

Index build

The first step of the pipeline was bowtie2-build, which builds a Bowtie index from the FASTA sequences set as a reference. Such index files is needed to align the reads to the reference at the mapping process [Langmead and Salzberg, 2012]. We used the instruction: `bowtie2-build reference_file.fasta bt2_index_base`.

Quality analysis

To start with the data analysis, quality of sequences was evaluated with the program FastQC v. 0.11.2 [Andrews, 2010]. The program was ran with a bash script that helped to submit all the files at one. Command line and parameters used in the script are shown in Appendix 1.

Mapping of sequenced reads

Reads were mapped with Bowtie 2 v 2.2.5 [Langmead and Salzberg, 2012] to a reference transcriptome, generated with a genome-guided de novo assembly in our group [Rougon-Cardoso *et al.,* 2016]. This software is an ultrafast and memory-efficient tool [Langmead and Salzberg, 2012]. Outputs were obtained in SAM (Sequence Alignment/Mapping) format, which is the standard format for alignments [Li *et al.*, 2009]. Bowtie 2 was also executed with a bash script. Command line and parameters used in the script are shown in Appendix 1.

Quantification of transcript abundance

To quantify the abundance of the mapped reads, we used eXpress [Roberts and Pachter, 2013]. The input files were the SAM-BAM files created in the previous step. This program was also executed with a bash script. Command line and parameters used in the script are shown in Appendix 1.

The eXpress results are obtained in a tab-delimited file. From these files, the “Effective Counts” and “length” columns were extracted and used to created a counts table. This table was created with an R [R Core Team, 2015] script from Dr. Cesaré Ovando. The script takes the transcript ids, and the library ids, to create a final tab-delimited table with the transcript id, the counts per library and the transcript length.

Differential expression analysis

With the created counts table we proceeded to perform the differential expression analysis in R using the Bioconductor package edgeR [Robinson *et al*., 2010]. This package performs a statistical analysis of expression data, taking into account variation between and among samples. Biological variation is modeled using a negative binomial distribution. The count table was filtered to have at least 4 counts per million in at least 9 of the 12 libraries. Then, data was normalized and the dispersion was estimated (across all genes). Then, the sample contrast was performed between hemolymph

inductions vs control at both time points. Differential transcripts were defined using a False Discovery Rate (FDR) cut-off of ≤0.1.

Differentially expressed proteins

Nematodes were induced with either hemolymph or intestine of *Galleria mellonella* for 4 hours. The induction protocol is the same as previously described in the "Nematode induction with insect hemolymph" section. Nematodes were then grinded in liquid nitrogen and used to extract total soluble proteins. Shotgun proteomics was done fractioning 200 µl of the total protein extract. Thirty µg of protein from each fraction were then analyzed by LC-MS / MS using a Thermo Scientific Q-Exactive Orbitrap MS spectrometer in conjunction with a Proxeon Easy-nLC II HPLC (Thermo Scientific) and a source Proxeon nanospray using a reverse phase column. The MS/MS spectra were acquired using the TOP15 method following the equipment manufacturer's instructions. All analyses were run in duplicates, including treatment samples and controls.

Analysis was first qualitative in which the proteins that were expressed in the samples and absent (below the detection level) in the controls, or vice versa, were detected. A second approach, using the same data, was a label-free quantitative analysis in which those proteins expressed only in the samples or the controls were excluded. Raw files of every fraction of the samples were processed using MaxQuant v 1.5.2.8 [Cox and Mann, 2008; Cox *et al*., 2014] for protein identification and quantification. For identification of proteins, a false discovery rate of 1% at the peptide and protein level was used. The average absolute mass deviation was 0.2 parts per million (p.p.m.). For protein quantification, we used intensity based absolute quantification, or iBAQ [Schwanhäusser *et al.*, 2011]. Proteins amounts were calculated as the sum of all peptide peak intensities divided by the number of theoretically observable tryptic peptides. Data analysis was done using Microsoft Office Excel and Perseus v.1.5.1.6 [Tyanova *et a*l., 2016]. Differential expression analysis was done using only proteins observed in the two replicates per condition, using t-test analyses and an FDR of 5%.

Results

For the first goal of estimating the impact that lifestyle has in the patterns of selection, we conducted genome-wide analysis of selection at both interspecific and intraspecific levels.

Positive selection in interspecific data

Orthologous proteins

For the interspecific analysis, orthologous genes were explored. The number of orthologues identified for the nematodes grouped by clades IV and V (Table 1) are shown in Table 4. The number of genes in the clade IV set, named C4N4, represent 9.5% of the 16,333 estimated genes for *S. carpocapsae* [Rougon-Cardoso *et al*., 2016], respectively. For clade V, set named C5N4 represent 6.91% of the 21,850 estimated genes for *C. briggsae* release WS244. These low number of orthologous genes identified correlates with the high divergence reported among nematodes species [Parkinson *et al.,* 2004].

Table 4. Number of orthologues identified

Dataset name	# of orthologous genes
C4N4	1552
C5N4	1510

Phylogeny reconstruction

Phylogenetic trees were obtained for the two datasets with four nematodes each, using the concatenated orthologous sequences for each dataset (Figure 3). The topologies are in agreement with previously published nematode phylogenies [Blaxter *et al*., 1998; Holterman *et al.,* 2006], and corroborate the phylogenetic relationships among the nematodes species on each clade (IV and V), as shown in previous work from our group [Rougon-Cardoso *et al.,* 2016].

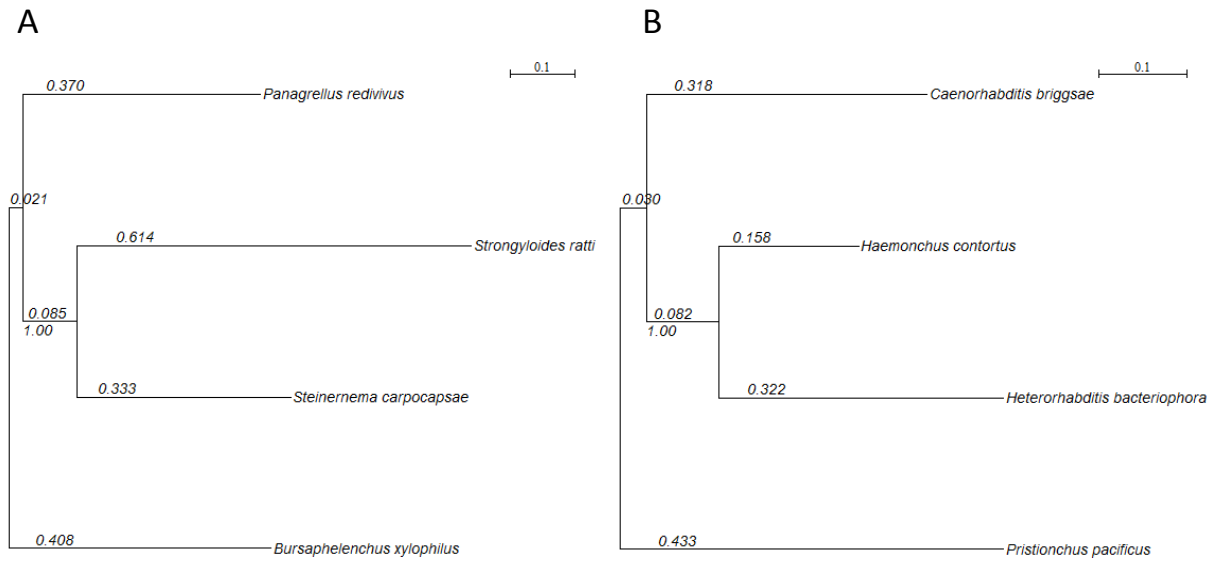


Figure 3. ML phylogenetic reconstructions for nematode of clade IV and V. A, clade IV; B, clade V. Values correspond to branch length. Trees are unrooted, and bootstrap values are shown under the branch leading to the animal parasites.

Scans of positive selection

In the C4N4 dataset, 124 (7.99%) of the tested genes had a different ω value among branches (LRT, $p < 0.05$). In C5N4, the number of genes showing different ω values among branches was 113 (7.48%) (LRT, $p < 0.05$). For this test, M0-M1, 5% of the genes are expected to be identified by chance. These results are indicative of episodic evolution, which is against the neutral expectation [Yang, 1998].

For the following analyses, we conducted four different LRTs for each dataset. We tested the branches of three different species and the branch leading to vertebrate parasites. *B. xylophilus* (clade IV) and *P. pacificus* (clade V) are basal in the phylogenies and their branches were not analyzed. This because *B. xylophilus* has a more complex lifestyle [Jones *et al*., 2008], making its comparison with the other nematodes more difficult to interpret. Especially with the clade V homologue branch, corresponding to *P. pacificus* which has a free-living lifestyle. Figures 4 and 5 illustrate the branches tested for each clade.

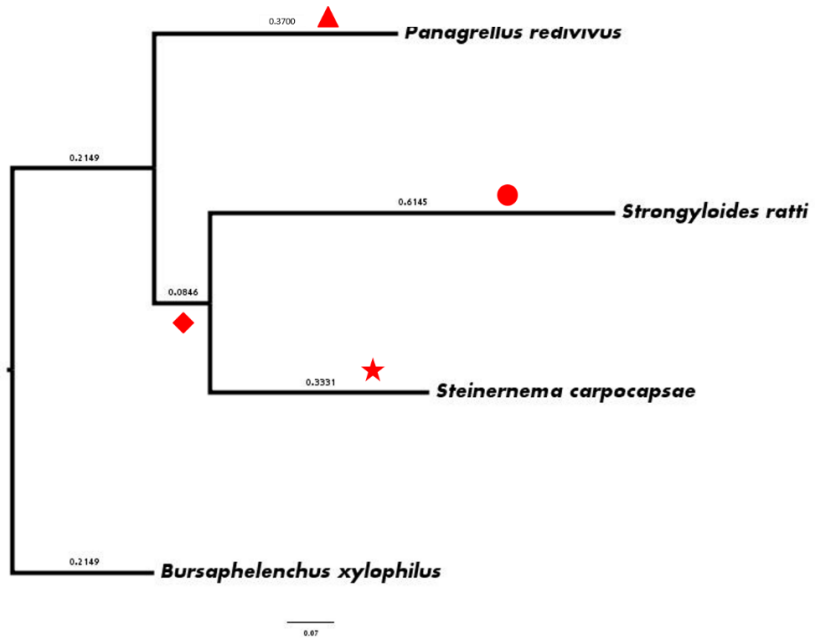


Figure 4. Clade IV phylogenetic tree showing the tested branches. Tested branches according to lifestyle; ✖, entomopathogenic; ●, vertebrate parasite; ♦animal parasite branch; ▲ , free-living. All tests are based on an unrooted phylogeny; the trees are rooted for display purposes only.

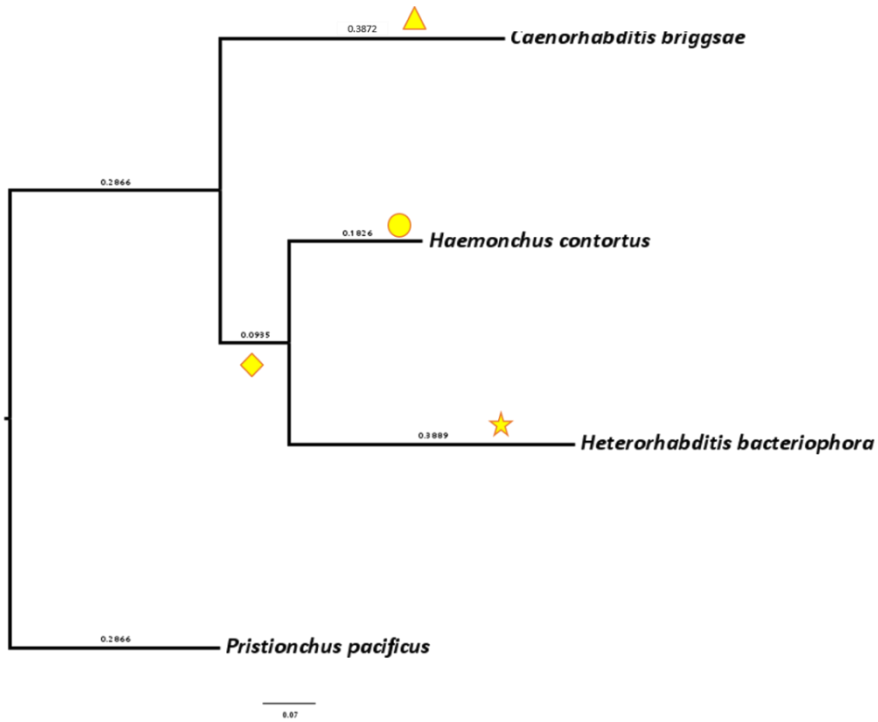


Figure 5. Clade V phylogenetic tree showing the tested branches. Tested branches according to lifestyle; ✖, entomopathogenic; ●, vertebrate parasite; ♦animal parasite branch; ▲ , free-living. All tests are based on an unrooted phylogeny; the trees are rooted for display purposes only.

When testing the one-ω model (model= 0) against a two-ω model that estimates one ω for a selected branch and one ω as the background ω for the remaining branches (model= 2) (See Table 2). We found a total of 79 genes, in all the species, in which the selected branch had an estimated ω > 1 and a significant different value to the background ω (Model 0-Model 2; LRT, $p < 0.05$). From these, 27 (1.73%) were from C4N4 and 52 (3.44%) from C5N4. In the C4N4 set, the *S. carpocapsae* branch had 9 genes (0.58 %) with these characteristics. For the rest of the branches tested, the percentage of genes ranged from 0.13-0.71 % in C4N4 and 0.46 -1.13% in C5N4 (Table 5).

To test if ω was significantly bigger than one in these cases, we contrasted a two-ratio model with ω fixed to one for the specified branch and a freely estimated ω for the remaining branches (model=2 – fix ω), against a two-ratio model as described above (model= 2). None of the genes tested had a significantly ω > 1 (LRT, $p > 0.05$) in any of the datasets. This means none of the genes are evolving under positive selection under the branch model test. However, we identified genes with a significant ω <1, i.e genes that are evolving under purifying selection. In the case of the C4N4 set, we found 1312 (84.54%) genes evolving under purifying selection in the *S. carpocapsae* branch; 1276 (82.22%) in the *Strongyloides ratti* branch; and 1345 (86.66%) in the *P. redivivus* branch. In the branch leading to vertebrate parasites (*S. carpocapsae* and *Strongyloides ratti*), we found that most of the genes, 1522 (98.07%) genes, were under purifying selection. In the C5N4 set we found 995 (65.89%) genes under purifying selection in the *Heterorhabditis bacteriophora* branch; 892 (59.07%) in the *Haemonchus contortus* branch; and 1042 (69.01%) in the *C. briggsae* branch. In the branch leading to vertebrate parasites (*Heterorhabditis bacteriophora* and *Haemonchus contortus*), we also found most of the genes, 1461 (96.75%), evolving under purifying selection. Genes evolving under neutrality showed a low proportion when testing branches corresponding to only one nematode species. The percentages of these genes ranged from 13.34-17.78% in C4N4 and 30.99-40.93% in C5N4. These results are summarized in Table 5.

It has been reported that branch models are conservative because positive selection often acts on one or a few amino acids, and averaging ω over sites results in a lack of power [Yang and Nielsen, 2002; Zhang *et al.,* 2005]. Therefore we also used branch-site models, created to detect positive selection affecting just a few sites along particular lineages [Yang and Nielsen, 2002; Zhang *et al.,* 2005]. We compared the branch-site model A, against the same model with the difference that ω_2 was fixed to 1 (model A1). For the C4N4 set, we found 74 (4.77%) genes with sites evolving under positive selection in the *S. carpocapsae* branch; 24 (1.55%) in the *Strongyloides ratti* branch; and 91 (5.86%) in the *Panagrellus redivivus* branch. In the branch leading to parasites (*S. carpocapsae* and *Strongyloides ratti*), we found 21 (1.35%) genes with sites under positive selection (Table 5). In the C5N4 set we found 61 (4.04%) genes with sites under positive selection in the *Heterorhabditis bacteriophora* branch; 55 (3.64%) in the *Haemonchus contortus* branch; and 87 (5.76%) in the *C. briggsae* branch. In the branch leading to parasites (*Heterorhabditis bacteriophora* and *Haemonchus contortus*), we found 20 (1.32%) genes with sites evolving under positive selection (Table 5). When analyzing nematodes based on their lifestyles, parasitic nematodes showed lower proportions of genes with sites under positive selection (214/6124 = 3.49%) than free-living nematodes (178/3062 = 5.81%), and the difference was highly significant (χ^2 test, $p<0.0000003$).

Table 5. Intraspecific analysis of positive selection in the four nematode datasets.

Dataset name	C4N4 Clade IV				C5N4 Clade V			
Orthologues analyzed	1552 9.50 % *				1510 6.91 % *			
Genes with ω significantly different among branches (LRT, p < 0.05)	124 7.99 %				113 7.48 %			
Foreground branch (ω1)	Sc	Sr	(ScSr)	Pr	Hb	Hc	(HbHc)	Cb
Branch model								
Genes with ω1 > 1 and significantly different than ω0 (LRT, p < 0.05)	9 0.58 %	11 0.71 %	2 0.13 %	5 0.32 %	17 1.13 %	11 0.73 %	7 0.46 %	17 1.13 %
Genes with ω1 significantly greater than 1 (LRT, p < 0.05)	0	0	0	0	0	0	0	0
Genes with ω1 significantly lower than 1 (LRT, p < 0.05)	1312 84.54%	1276 82.22%	1522 98.07%	1345 86.66%	995 65.89%	892 59.07%	1461 95.75%	1042 69.01%
Genes under neutrality	240 15.46%	276 17.78%	30 1.93%	207 13.34%	515 34.11%	618 40.93%	49 3.25%	468 30.99%
Branch-site model								
Genes with sites under positive selection (ω > 1) (LRT, p < 0.05)	74 4.77 %	24 1.55 %	21 1.3 % 5	91 5.86 %	61 4.04 %	55 3.64 %	20 1.32 %	87 5.76 %
Average proportion of sites under positive selection per gene (standard deviation)	4.94 % (0.053)	6.31 % (0.117)	3.58 % (0.045)	5.16 % (0.054)	9.88 % (0.140)	5.81 % (0.068)	8.74 % (0.155)	7.90 % (0.092)

Percentages are from the total of genes tested for each set unless stated. * Percentage of genes in relation to the total genes estimated for *S. carpocapsa*e in clade IV and *C. briggsae* in clade V. Sc, *S. carpocapsae*; Sr, *Strongyloides ratti*; Pr, *Panagrellus redivivus;* Hb, *Heterorhabditis bacteriophora;* Hc, *Haemonchus contortus*; Cb, *Caenorhabditis briggsae*. The tests for positive selected genes were significant (LRT, p<0.05)

Functional Analysis of Positively Selected Genes and Over-Represented Categories

All genes were associated to functional categories from Gene Ontology (GO). In genes, with positively selected sites at least one GO category was identified for 180 (85.71%) genes in the set C4N4, 183 (82.06%) genes in the set C5N4. Functional annotation of *S. carpocapsae* genes with sites under positive selection is shown in Figure 6 at a depth level 4. GO is structured as a graph, terms would appear at different 'levels' if different paths were followed through the graph. Even though GO terms do not occupy strict fixed levels of hierarchy, we obtained a rough classification on Blast2GO.

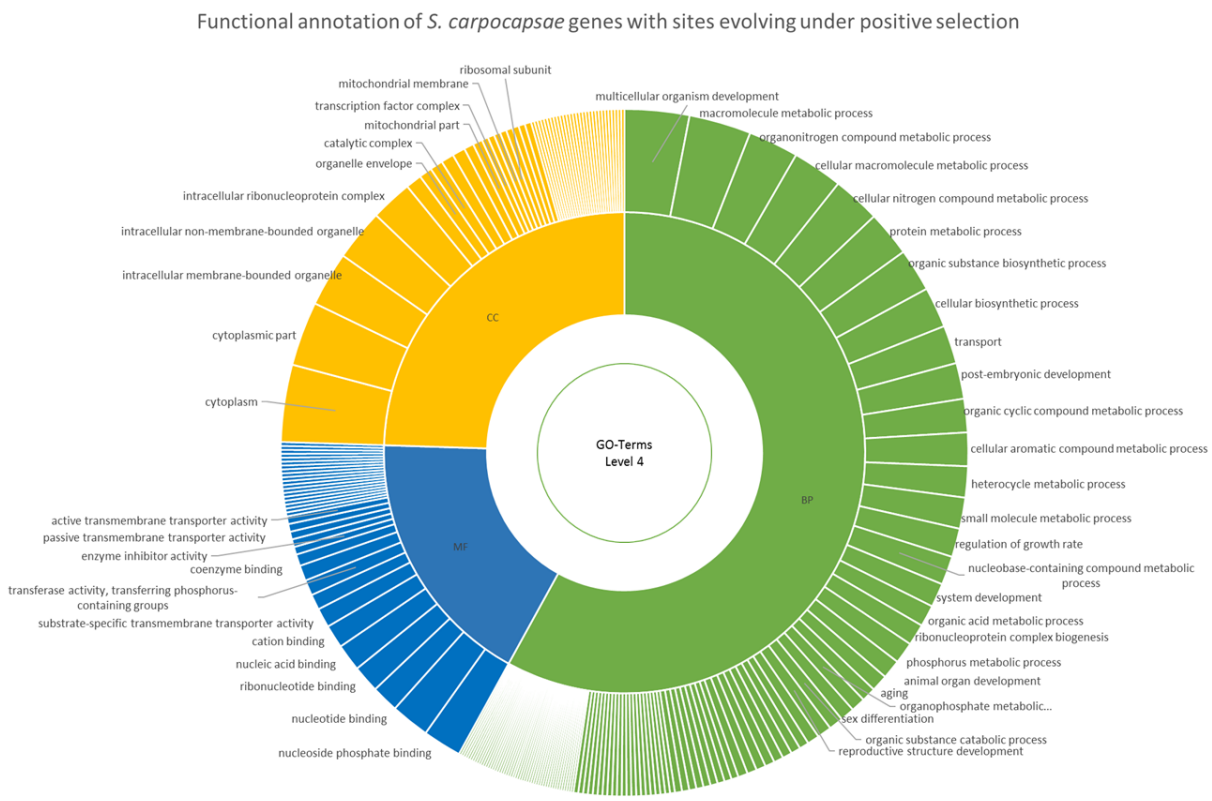


Figure 6. Functional annotation of *S. carpocapsae* genes with sites evolving under positive selection. GO Distribution by Level (4) of *S. carpocapsae* genes with sites evolving under positive selection. Biological process are shown in green, molecular function in blue, and cellular component in yellow.

When looking for over represented functional categories, we found that from the 210 protein-coding genes in clade IV, 26 were significantly over-represented (enriched) unique GO terms (Appendix 2). From these, 19 were enriched only in *S. carpocapsae*, one in *Strongyloides ratti*, five in *P. redivivus*, and only one was shared among the three species. In the 223 total protein-coding genes of clade V, we found 28 significantly enriched unique GO terms (Appendix 3), six of which were unique to *Heterorhabditis bacteriophora*, 18 were found only in *Haemonchus contortus*, and four were shared in both nematodes. Not enriched term was found among the *C. briggsae* genes with signs of positive selection. Only one GO term resulted to be enriched in both clades IV and V (GO:0005198 structural molecule activity); therefore, there were 53 unique enriched terms considering both clades. In general, in parasitic nematodes the total number of enriched terms (47) was notoriously high in relation to free-living nematodes (5), even

considering that the number of genes with signatures of positive selection was lower in parasitic nematodes. Most of the enriched terms in *S. carpocapsae* are related to the immune response or antimicrobial peptide production. Only one term related to mitochondrial function was found in *S. carpocapsae*, however, in the parasite *Haemonchus contortus,* an important number of terms related to mitochondria were observed (see Appendices 2 and 3).

Selection in intraspecific data

Species and sequences

Strains from two nematode species, *S. carpocapsae* and *C. briggsae* (Denver *et al.,* 2012; Thomas *et al.*, 2015), were used for the intraspecific analysis. Strains details are shown previously in Table 3 in the Material and Methods section. We selected *C. briggsae* as a model of free-living nematode for comparison because it was the only free-living species with several available genomes from different strains [Thomas *et al.,* 2015]. *C. briggsae* sequence data was downloaded from the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra).

For this study, the genomes of four *S. carpocapsae* strains were sequenced. DNA isolation and sequencing yields are shown in Table 6. The quality of the sequences was performed with FastQC in a non-interactive mode.

Table 6. DNA concentration and quality and sequencing yield of *S. carpocapsae* strains

Sample	Concentration ng/µl	260/230	260/280	Number of reads
Sc ALL	1639	2.19	2.18	34,695,646
Sc Az20	809.4	2.16	2.13	34,030,662
Sc Az154	1,444	2.27	2.20	34,344,025
Sc Az157	997.1	2.09	2.13	38,407,029

Mapping and SNP determination

Reads were mapped to the reference strain for each nematode species, (*S. carpocapsae* (Breton) and *C. briggsae* (AF16). The percentage of mapped reads ranged from 96.5% to 100%, results per strain are shown in Table 7.

Table 7. Percentage of mapped reads to references, *S. carpocapsae* (Breton) and *C. briggsae* (AF16)

Sample	Percentage of reads mapped
Sc ALL	96.72%
Sc Az20	96.60%
Sc Az154	96.50%
Sc Az157	96.69%
Cb JU1348	97.33%
Cb VX0034	100.00%
Cb QR25	96.68%
Cb ED3101	97.29%

The mapped reads were used in the SNP variant search and then used to construct haplotype map files. An example of the structure and content of these files is shown below.

Scaffold/Chromosome	Base	Reference	Strain 1	Strain 2	Strain 3	Strain 4
scaffold00001	5	T	.	C	.	.
scaffold00001	11	T	C	.	.	.
scaffold00001	14	C	.	G	T	T
scaffold00001	22	T	.	C	.	C
scaffold00001	24	A	.	G	.	G

Selection scans and statistical tests

We conducted tests of neutrality with Tajima's D statistic in non-overlapping sliding windows of 1000 bp across the genome, including both genic and intergenic regions. In windows without variation (π=0) the estimation of Tajima's D is not possible. This windows were called invariable and were left aside for most of the analysis.

Tajima's D statistic is sensitive to demographic processes that might produce signals that can be confused with signatures of selection [Nielsen 2005]. To account for the demographic effects that could be acting upon the species analyzed, we used a program developed in our group that uses the observed distribution to calculate corrected confidence limits. This correction method was studied by Schmidt and Pool. They used the observed distribution to correct the confidence limits in simulated distributions [Schmidt and Pool 2002]. Excluding demographic processes, significant positive values of Tajima's D are indicative of balancing selection, whereas significant negative values are indicative of directional selection [Tajima 1989]. Directional selection can be due to either positive or negative selection [Nielsen 2005].

From the *S. carpocapsae* assembled genome, we analyzed 84,767 windows, from which 69,694 had enough coverage to condcut the analysis (Table 8). From these, 400 windows (0.47%) were invariable. The remaining 69,294 windows, which included 14,994 protein-coding genes (representing 91.84% of the protein-coding genome) were used to build the observed distribution of Tajima's D values shown in Figure 7.

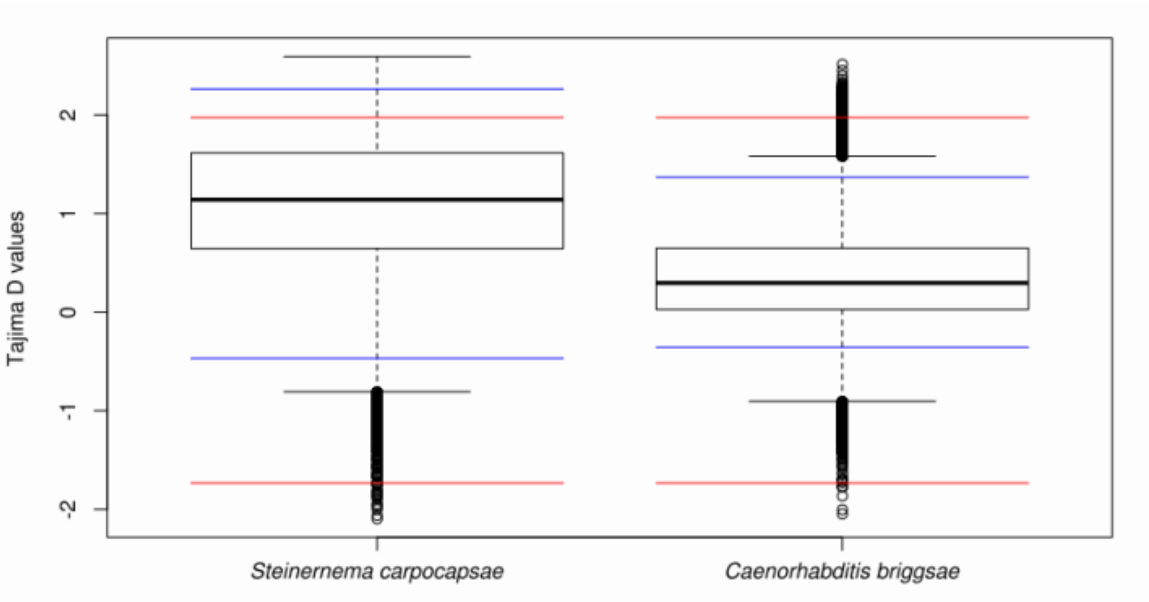


Figure 7. *S.carpocapsae* and *C. briggsae* Tajima's D genome distribution. Box plots of the distribution of Tajima's D values estimated in windows of 1,000 bp across the *S. carpocapsae* and *C. briggsae* genomes. Red lines correspond to the maximum and minimum confidence limits according to [Tajima 1989]. Blue lines correspond to the maximum and minimum corrected confidence limits from the observed distribution (upper and lower 2%).

Tajima's D values ranged from -2.105 to 2.592. The top 2% of positive values in the distribution were considered as significant, and included 1,375 windows comprising 228 protein-coding genes (1.52% of the genes analyzed, see Appendix 4). The top 2% of negative values included 1,384 windows comprising 302 protein-coding genes (2.01%). The total number of genes with significant D values was 530 (3.53% of the genes analyzed, see Appendix 4). The confidence limits were -0.468 for negative and 2.264 for positive values, which differed from the uncorrected limits obtained with the Tajima's beta distribution (-1.733 and 1.975). The number of windows with significant values according to the theoretical distribution are shown in Table 8 for comparison.

Table 8. Intraspecific analysis of selection in *S. carpocapsae* and *C. briggsae*

Species	*S. carpocapsae*	*C. briggsae*
Number of windows	84,767	108,421
Windows with coverage > 90%	69,694	65,933
Invariable windows	400	21
Windows with Tajima's D values	69,294	65,912
Genes covered (>50%) in windows with D values	14,994	15,473
Tajima's D range of values	-2.105 — 2.592	-2.057 — 2.452
Tajima´s D average	1.097	0.355
Confidence limits according to theoretical distribution [28]	-1.733 — 1.975	-1.733 — 1.975
Theoretical significant windows	6522 (6500 positive, 22 negative)	80 (74 positive, 6 negative)
Confidence limits obtained from the observed distribution	-0.468 — 2.264	-0.355 — 1.369
Significant windows according to the observed distribution*	2,759 (1,375 positive, 1,384 negative)	2,624 (1,313 positive, 1,311 negative)
π values range	0.00035 — 0.075	0.00035 — 0.045
Average π value	0.006700	0.007763
Θ values range	0.35349 - 61.5065	0.3535 - 49.1345
Average Θ values	5.369212	7.166656

* Windows falling in the upper (positive) and lower (negative) 2% of the observed distribution.

Tajima's D significant values were plotted with the π values in 10kb regions to see the surrounding windows behavior. Figure 8 shows an example of genes with negative and positive significant Tajima's D. Panel A in this figure shows the orthologous gene of *Trichinella spiralis* STK11, coding for a serine/threonine-protein kinase (g2901/ uniprot A0A0V1AZR9) associated to ATP binding, which presented the most negative D value of the distribution. Panel B corresponds to the orthologous gene of *C. elegans* nuo-6, coding for a subunit of the mitochondrial NADH dehydrogenase (ubiquinone) complex (complex I) (g6196|uniprot Q23098), which presented the most positive D value.

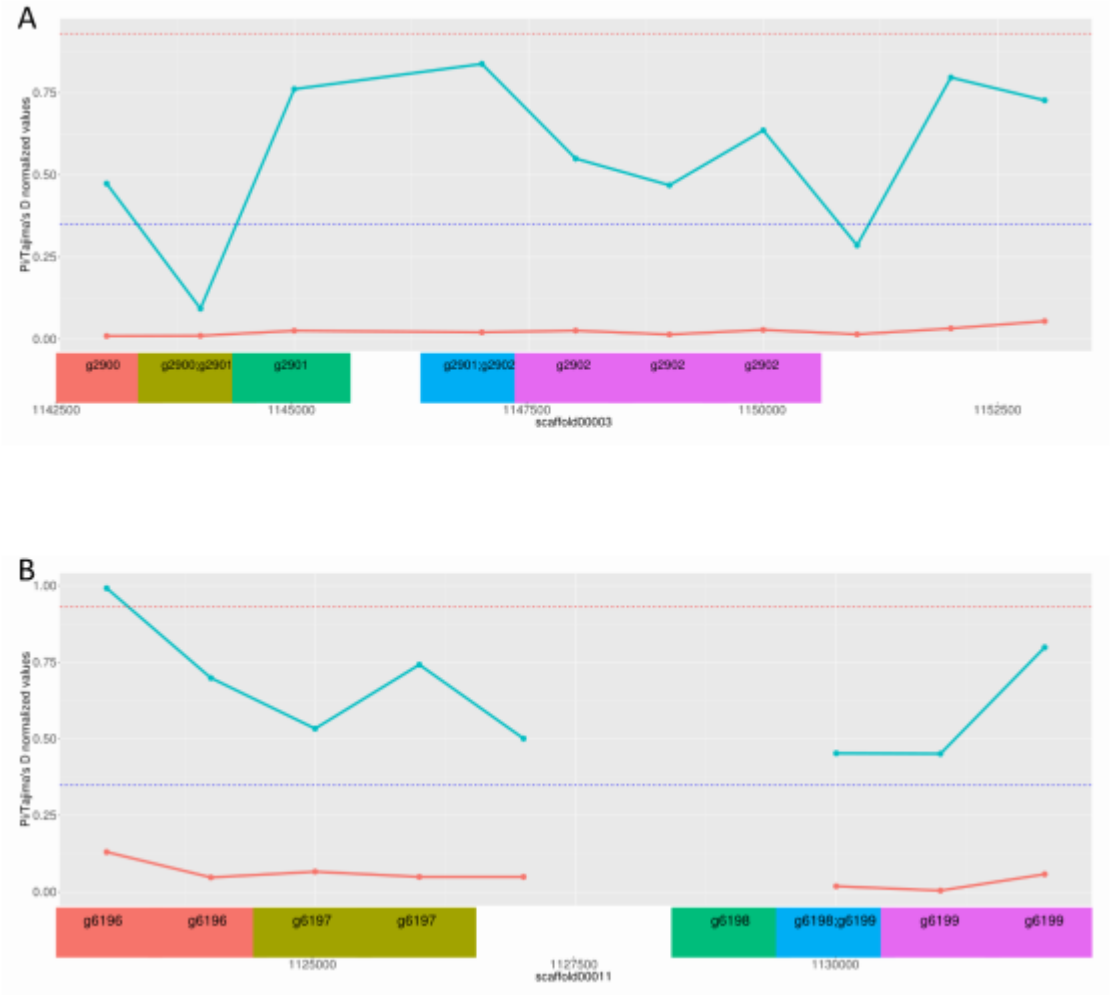


Figure 8. Tajima's D and π values in a 10kb genomic region. Tajima's D values in solid teal line and π values in solid red line. Each point represents the value calculated for each of the 1,000 bp windows analyzed across the genome. Both Tajima's D and π values are normalized for display purposes. Dashed lines indicate the corrected confidence lower (blue) and upper (red) limits. Grid corresponds to the gene coordinates in the scaffolds along the X axis. Genes are represented by the color boxes, with each gene indicated with a different color.

For *C. briggsae* we analyzed 108,421 windows, from which 65,933 had enough coverage to conduct the analysis (Table 8). From these, 21 windows (0.02%) were invariable. The remaining 65,912 windows, which included 15,473 protein-coding genes (representing 66.52% of the protein-coding genome) were used to build the observed distribution of Tajima's D values (Figure 7). D values ranged from -2.057 to 2.452. The top 2% of positive values in the distribution included 1,313 windows comprising 247 protein-coding genes (1.6% of the genes analyzed). The top 2% of negative values included 1,311 windows comprising 240 protein coding genes (1.55%). The total number of genes with significant D values was 487 (3.15% of the genes analyzed). The confidence limits (96%) from the observed data distribution were -0.355 and 1.369 (with -1.733 and 1.975 (95%) for the theoretical beta distribution). The comparison with results using the theoretical distribution are shown in Table 8.

The average of Tajima's D values in *S. carpocapsae* was bigger (1.097) and significantly different than the average in *C. briggsae* (0.355) (W test, $p=2.2x10^{-16}$) (Figure 7). This was not due to a simple displacement of the distribution of one species in relation to the other but to a clear change in the shape of the distribution (K-S test, $p=2.2x10^{-16}$). Figure 9 shows the distribution of Tajima's D values different between these species, with the distribution of *S. carpocapsae* more skewed to positive values.

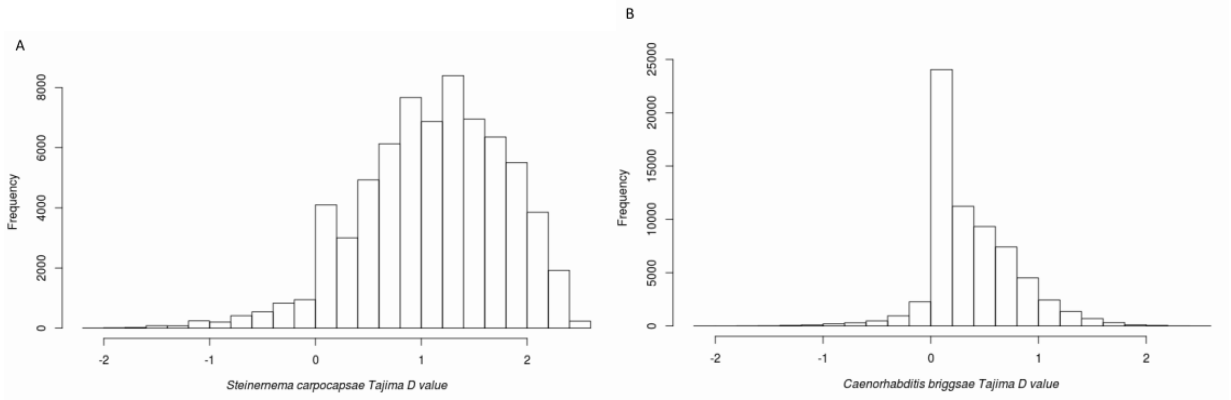


Figure 9. Distribution of Tajima's D values estimated in sliding windows of 1,000 bp across the genomes. A) *S. carpocapsae*; B) *C. briggsae*. D values were non-continuous (i.e. were distributed in discrete intervals).

Functional annotations of *S. carpocapsae* genes in windows with significant Tajima's D value are shown in Figure 10. In *S. carpocapsae*, there were no enriched GO terms in the genes with significant values, either positive or negative.

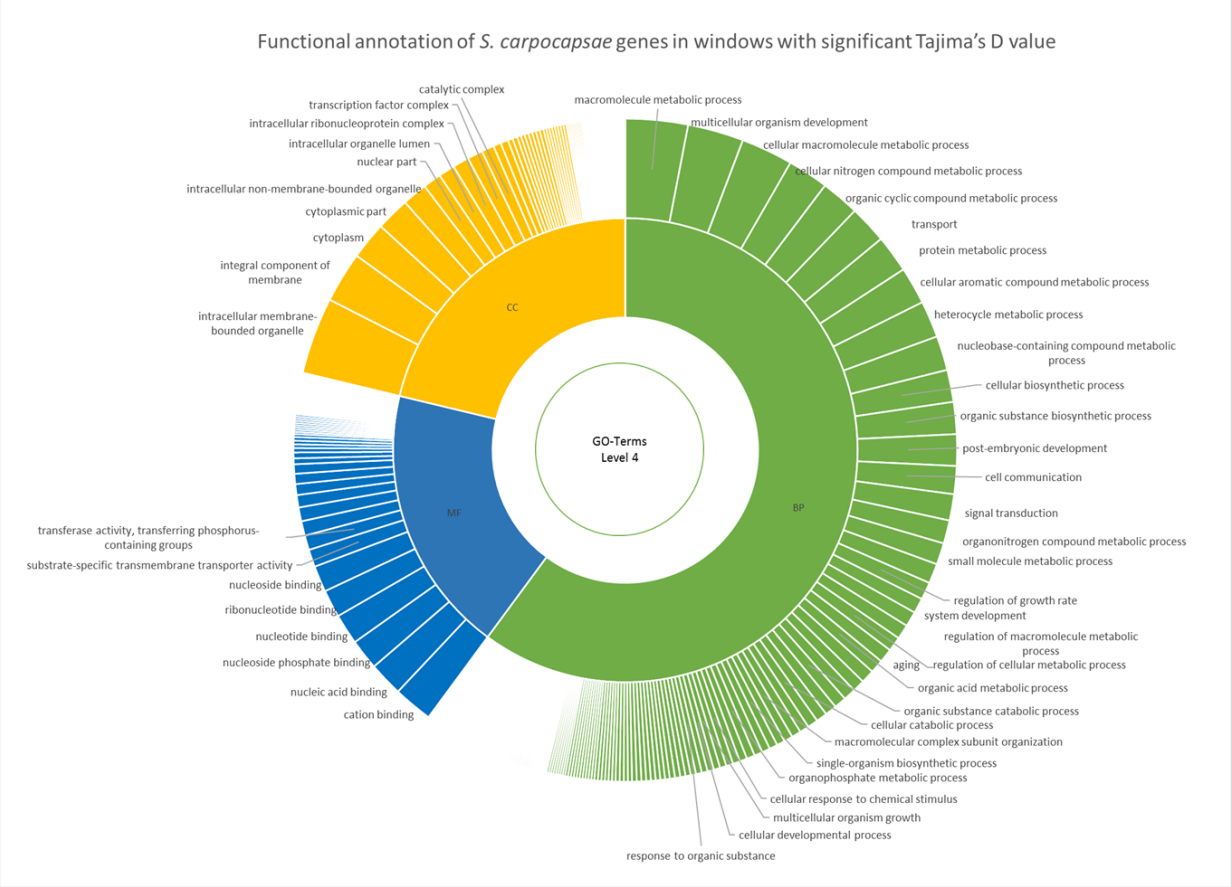


Figure 10. Functional annotation of *S. carpocapsae* genes in windows with significant Tajima's D value. GO distribution by level (4) of *S. carpocapsae* protein coding genes included in windows with significant Tajima's D values. Biological process (BP) in green, molecular function (MF) in blue, and cellular component (CC) in yellow of the genes in windows with significant Tajima's D value. GO analysis was performed using B2GO.

Differentially expressed genes due to the interaction with insect tissue

Infection kinetics

To have a better approximation to when IJs reach the intestines and hemocoel, infected *G. mellonella* larvae were dissected, and nematodes present in each insect compartment were counted. Each experiment was repeated at least 2 times. We found almost no presence of nematodes at 3 hours post infection. After 4 hours post infection, nematodes were present mainly in the intestine (Figure 11). After this time, we found nematodes in both tissues, with a prevalence in the larval intestine.

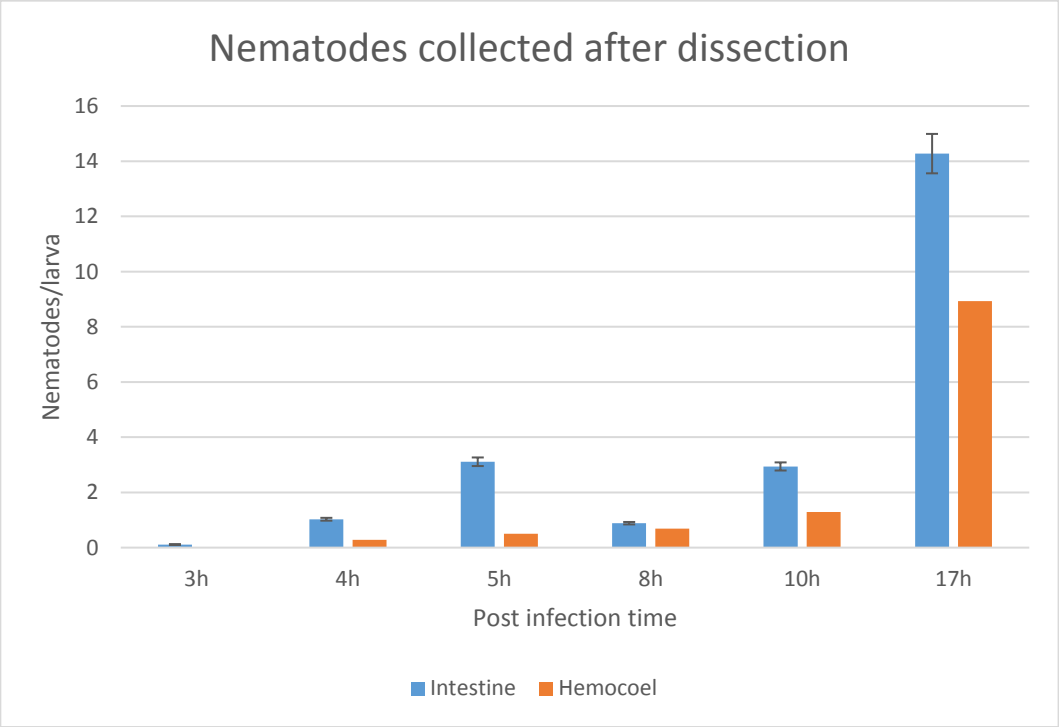


Figure 11. Nematodes collected from *G. mellonella* tissues at different time points post-infection.

Results from these experiments allowed us to estimate the time when IJs reach the hemocoel and to propose a model for the infection process at early stages (Figure 12).

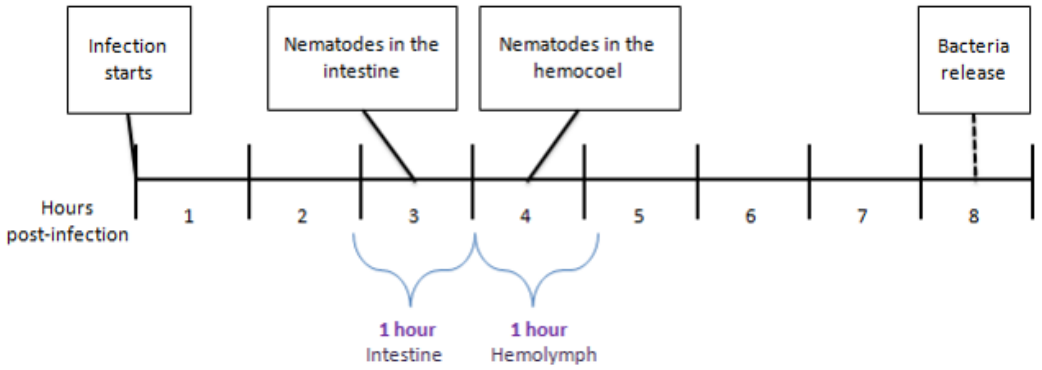


Figure 12. Nematode infection model. Nematodes are present in the intestine at three hours post infection and in the hemocoel at four hours post infection. Incubation with insect tissues was designed to simulate the arrival to these tissues.

Nematode induction and libraries preparation

We used the proposed infection model to simulate the infection process. We induced *S. carpocapsae* IJs for 1 or 2 hours with insect (*G. mellonella*) hemolymph. Then washed and grinded as stated in the methods. We performed three biological replicates of each experiment. The replicates were made in consecutive dates and with extreme care to maintain the nematodes frozen. Table 9 shows the library details (induction conditions, times, and RNA extractions yields).

Table 9. RNASeq libraries details

	Sample	Condition	Incubation time	replicate	Induction date	260/280	260/230	ng/µL	Library ID
1	C1.1	Control	1 hour	1	11.03.2015	2.09	1.47	1403.8	RM1TS1SS25.13
2	C1.2	Control	1 hour	2	12.03.2015	2.05	1.16	920.4	RM1TS1SS26.01
3	C1.4	Control	1 hour	3	19.03.2015	1.93	0.78	394.2	RM1TS1SS27.02
4	C2.1	Control	2 hours	1	11.03.2015	2.07	1.44	847.1	RM1TS1SS28.03
5	C2.2	Control	2 hours	2	12.03.2015	2.1	1.57	1011.3	RM1TS1SS29.04
6	C2.4	Control	2 hours	3	19.03.2015	2.1	0.91	906.6	RM1TS1SS30.05
7	H1.1	Hemolymph	1 hour	1	11.03.2015	2.11	1.54	1297.1	RM1TS1SS31.06
8	H1.2	Hemolymph	1 hour	2	12.03.2015	2.08	1.38	883.2	RM1TS1SS32.07
9	H1.4	Hemolymph	1 hour	3	19.03.2015	2.06	0.54	848.9	RM1TS1SS33.08
10	H2.1	Hemolymph	2 hours	1	11.03.2015	2.04	1.09	929.6	RM1TS1SS34.09
11	H2.2	Hemolymph	2 hours	2	12.03.2015	2.1	0.76	819.5	RM1TS1SS35.10
12	H2.4	Hemolymph	2 hours	3	19.03.2015	2.09	0.82	1401	RM1TS1SS36.11

RNA sequencing and quality control

We had several failed sequencing runs maybe due to initial delays in the sequencing runs that ultimately led to libraries degradation (Table 10). Sequences include two runs of MiSeq and four on a HiSeq 2500. The MiSeq runs yielded an amount of reads lower, than needed for the differential expression analysis, making the change of sequencing platform. The third and last round was done in one lane of a HiSeq 3000 platform. The lane contained the 12 Paired-end TruSeq RNA-Seq constructed from three replicas of each of the induction times and its controls. Sequencing yields are shown in Table 11, including reads per library, average quality, and percentage of duplicate reads.

Table 10. Sequencing runs preformed during the project

Data type	Dates	Runs	Number of reads	Format	Duplicate	Average Mapping
MiSeq	Jan 2013 / Feb 2013	2	600,000	2x250		67%
HiSeq 2500	Jan 2014/Jul 2014	3	2.5 Million	2x100	30%	65%
	Aug 2014/Sep 2014	1	2.5 Million		50%	60%
HiSeq 3000	Sep 2015	1	50 Million/library		50%	60%

Table 11. Sequenced outputs and quality statistics per library

Sample	Files	Reads per file	Total reads per library	Average quality	Percentage of duplicates
C1.1	RM1TS1SS25-13_S78_L008_R1_001_fastqc RM1TS1SS25-13_S78_L008_R2_001_fastqc	23,792,040	47,584,080	40	61.9
				36	60.69
C1.2	RM1TS1SS26-01_S79_L008_R1_001_fastqc RM1TS1SS26-01_S79_L008_R2_001_fastqc	28,413,912	56,827,824	40	61.05
				35	58.77
C1.4	RM1TS1SS27-02_S80_L008_R1_001_fastqc RM1TS1SS27-02_S80_L008_R2_001_fastqc	32,904,662	65,809,324	38	65.19
				35	65.41
C2.1	RM1TS1SS28-03_S81_L008_R1_001_fastqc RM1TS1SS28-03_S81_L008_R2_001_fastqc	31,797,124	63,594,248	38	61.49
				36	60.6
C2.2	RM1TS1SS29-04_S82_L008_R1_001_fastqc RM1TS1SS29-04_S82_L008_R2_001_fastqc	33,171,058	66,342,116	39	62.12
				35	61.62
C2.4	RM1TS1SS30-05_S83_L008_R1_001_fastqc RM1TS1SS30-05_S83_L008_R2_001_fastqc	27,375,488	54,750,976	37	60.99
				37	59.88
H1.1	RM1TS1SS31-06_S84_L008_R1_001_fastqc RM1TS1SS31-06_S84_L008_R2_001_fastqc	32,703,765	65,407,530	40	61.94
				37	61.48
H1.2	RM1TS1SS32-07_S85_L008_R1_001_fastqc RM1TS1SS32-07_S85_L008_R2_001_fastqc	29,536,434	59,072,868	39	59.63
				37	58.09
H1.4	RM1TS1SS33-08_S86_L008_R1_001_fastqc RM1TS1SS33-08_S86_L008_R2_001_fastqc	26,653,889	53,307,778	40	60.65
				37	58.6
H2.1	RM1TS1SS34-09_S87_L008_R1_001_fastqc RM1TS1SS34-09_S87_L008_R2_001_fastqc	28,626,565	57,253,130	40	61.18
				37	60.73
H2.2	RM1TS1SS35-10_S88_L008_R1_001_fastqc RM1TS1SS35-10_S88_L008_R2_001_fastqc	31,824,896	63,649,792	40	61.88
				37	61.45
H2.4	RM1TS1SS36-11_S89_L008_R1_001_fastqc RM1TS1SS36-11_S89_L008_R2_001_fastqc	28,821,477	57,642,954	40	59.73
				38	58.94

The quality of the sequences was assessed with FastQC in a non-interactive mode. This mode was integrated into the pipeline used for the analysis, which allowed us to process all the files with a bash script. Figure 13 shows an example of the quality graphs we obtained in average per base for each read. The quality of the bases appeared to be in the range of 30 to 40 on the Phred scale. This range corresponds to a 99.9-99.99% of accuracy, indicating a very low error rate. All the other libraries showed similar quality values, most of them between the same range, with most of the R1 reads close to a Phred score of 40, and the R2 reads ranging from 35-37 of Phred scores (see table 11).

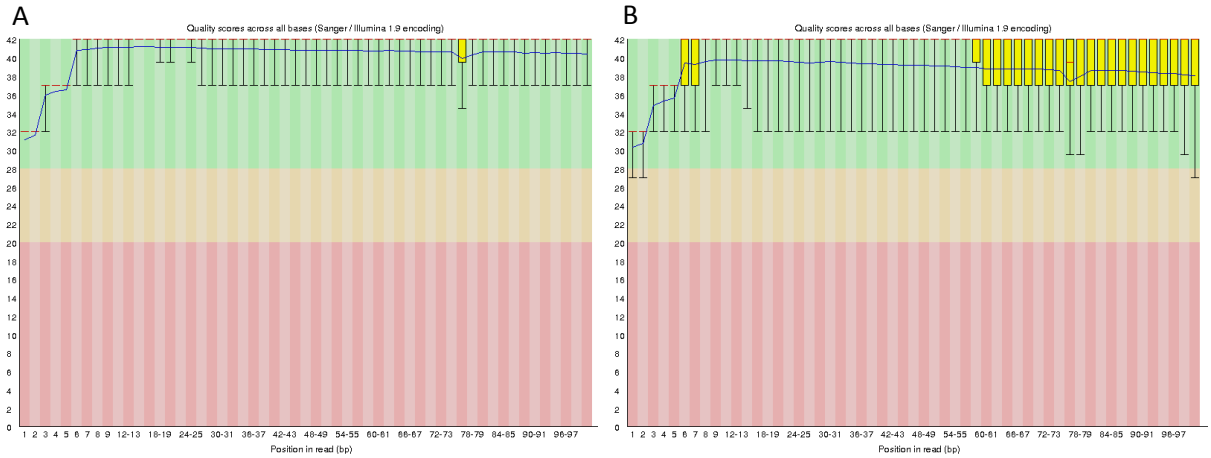


Figure 13. Example of quality scores across the reads. Reads from PE library C2.1 (RM1TS1SS28.03). A, corresponds to read 1 file; B, corresponds to read 2 file.

Mapping and quantification of transcript abundance

After the quality control step, reads were mapped to annotated transcripts generated in our group (Rougon-Cardoso *et al.,* 2016). Obtaining a slightly higher percent than if mapping to the genome. The percentage of mapped reads ranged from 55.8% to 65.1%, results per library are shown in Table 12.

Once the mapping was done, we quantified the abundance of transcripts and constructed a count table file with eXpress, using the effective counts and transcript length. The effective counts correspond to the expected number of reads from the sequencing experiment if sequencing and length biases did not exist [Roberts and Pachter, 2013]. These rounded counts are widely used for differential expression analysis, as they can be directly compared across experiments after taking into account sequencing depth [Robinson *et al*., 2010].

Table 12. Mapping results for each RNAseq library of induced IJs.

Sample	Total reads per library	Total bp	Duplicates	Percentage of reads mapped
C1.1	47,584,080	4,758,408,000	61.9%	55.81%
			60.69%	
C1.2	56,827,824	5,682,782,400	61.05%	57.91%
			58.77%	
C1.4	65,809,324	6,580,932,400	65.19%	50.56%
			65.41%	
C2.1	63,594,248	6,359,424,800	61.49%	63.87%
			60.6%	
C2.2	66,342,116	6,634,211,600	62.12%	65.08%
			61.62%	
C2.4	54,750,976	5,475,097,600	60.99%	60.12%
			59.88%	
H1.1	65,407,530	6,540,753,000	61.94%	64.57%
			61.48%	
H1.2	59,072,868	5,907,286,800	59.63%	63.14%
			58.09%	
H1.4	53,307,778	5,330,777,800	60.65%	57.25%
			58.6%	
H2.1	57,253,130	5,725,313,000	61.18%	59.49%
			60.73%	
H2.2	63,649,792	6,364,979,200	61.88%	62.58%
			61.45%	
H2.4	57,642,954	5,764,295,400	59.73%	62.51%
			58.94%	

Differential expression analysis

To search for consistency among replicates, we generated multidimensional scaling (MDS) plots. Figure 14 shows the MDS plots for each of the induction times. In these plots, the distances on the plot correspond to leading log-fold-changes between the RNA samples. The leading log-fold-change is the average (root-mean-square) of the largest absolute log-fold-changes between each pair of samples [Robinson *et al*., 2010]. Neither of the plots showed a clear separation between the two conditions (induction and control). Also, in both time points one of the replicates has a bigger distance than the rest of the replicates.

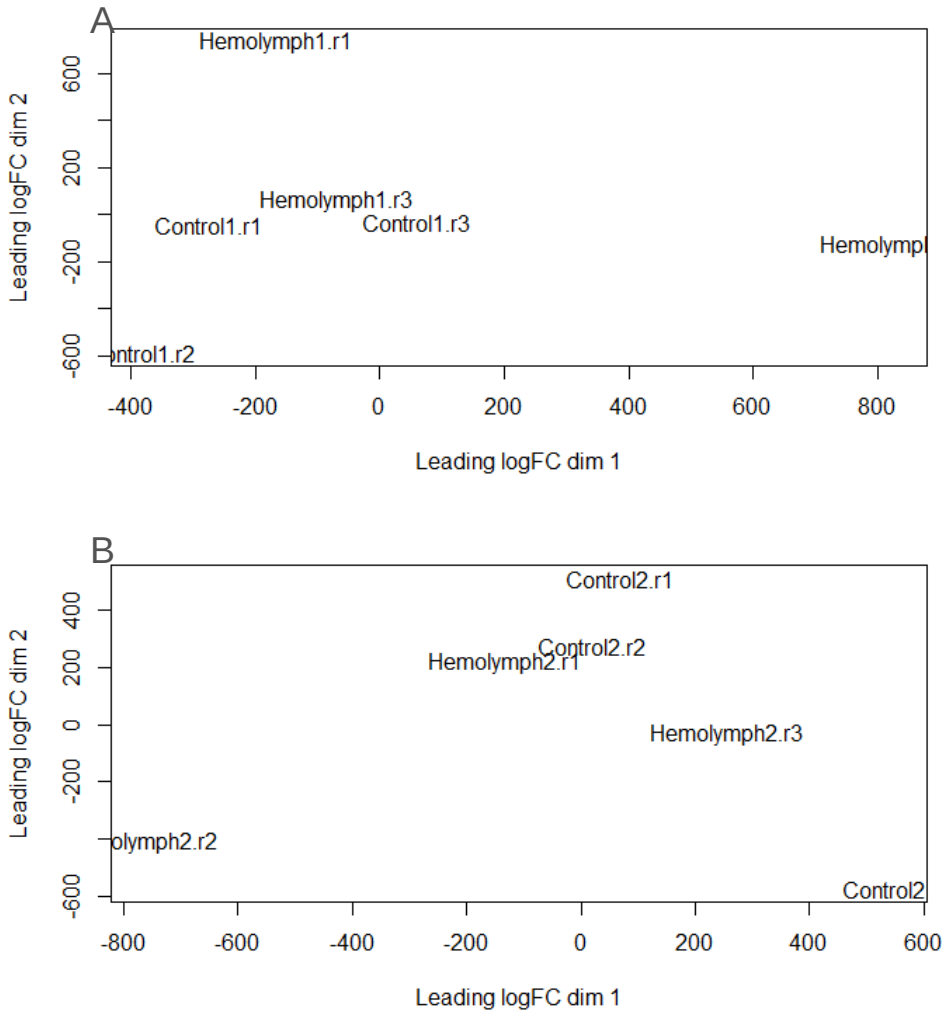


Figure 14. Multidimensional scaling plots of the IJs expression profiles. Plotting correspond to the counts per million (cpm) for A, one hour induction; B, two hour induction.

We estimated the biological coefficient of variation (BCV), which is the coefficient with which the abundance of the gene varies between replicate RNA samples. It is often assumed that all genes will have the same dispersion and that the technical coefficient of variation will decreases as the size of the counts increases [Robinson *et al*., 2010; McCarthy *et al.,* 2012]. However, the BCV is variable and it could reflect the differences in abundance of each gene between replicate RNA samples. Reported typical values for the common BCV for datasets from well-controlled experiments are: 0.4 for human data,

0.1 for data on genetically identical model organisms and 0.01 for technical replicates [Robinson *et al*., 2010]. Surprisingly, our data had a BCV of 2.564278, which indicates a high difference between replicates. Correlation analysis and a multi scatter plot of the samples are shown in Appendix 5.

Finally, when contrasting hemolymph inductions vs control at both time points we found very few differentially expressed genes as consequence of the high variation among samples (see Table 13).

Table 13. *S. carpocapsae*'s differentially expressed genes during insect tissue interaction

Time		Seq. Name	Seq. Description	GO Terms
1h	up	g3597	c-type lectin domain-containing protein 160	F:carbohydrate binding; F:protein binding
	down	g13594	tyrosine-protein kinase fer	P:response to stimulus; F:protein kinase activity; F:protein binding; C:intracellular part; F:ATP binding; P:protein phosphorylation
	down	g14449	cytosolic 10-formyltetrahydrofolate dehydrogenase	F:oxidoreductase activity, acting on the CH-NH group of donors, NAD or NADP as acceptor; F:binding; F:substrate-specific transporter activity; P:10-formyltetrahydrofolate metabolic process; P:one-carbon metabolic process; C:intracellular part; F:methyltransferase activity; P:biosynthetic process; P:oxidation-reduction process
2h	Up	g2033	YEATS domain containing protein 4	P:vulval development; P:posttranscriptional gene silencing by RNA; P:embryo development; C:nucleus
	Up	g12790	protein spec3-	----
	Up	g13564	phosphatidylinositol n-acetylglucosaminyltransferase subunit c	P:GPI anchor biosynthetic process; C:integral to membrane; F:phosphatidylinositol N-acetylglucosaminyltransferase activity
	Up	g14400	---NA---	----
	Up	g15901	cytochrome c oxidase subunit iv	P:embryo development; P:multicellular organismal aging; P:larval development; F:cytochrome-c oxidase activity; P:mitochondrial electron transport, cytochrome c to oxygen; P:proton transport; C:respiratory chain complex IV
	Down	g5270	---NA---	----
	Down	g9446	---NA---	----
	Down	g10193	protein isoform d	----
	Down	g14642	---NA---	----
	Down	g15104	hypothetical protein WUBG_05660	----

We found three genes in the 1 hour induction, one up-regulated and two down-regulated. The only up regulated gene that we detected was the one coding for a c-type lectin domain-containing protein 160, which is an orthologue of the *C. elegans* gene clec-160, which in turn is an orthologue of human FCER2 (Fc fragment of IgE, low affinity II, receptor for (CD23)) [Spencer *et al.,* 2011]. This gene has essential roles in the regulation of IgE production and in the differentiation of B-cells (it is a B-cell-specific antigen). The other two differentially expressed genes were down-regulated. The one coding for a tyrosine-protein kinase fer, had a *C. remanei* homologue, CRE29498, which based on protein domain information is involved in protein phosphorylation and has a tyrosine kinase activity. Finally, the gene coding for the cytosolic 10-formyltetrahydrofolate dehydrogenase protein is an orthologue of the *C. elegans* gene alh-3, which encodes an aldehyde dehydrogenase predicted to be mitochondrial.

In the two hour induction experiment we identified 10 differentially expressed genes, five up-regulated and 5 down-regulated. All the five up-regulated genes had orthologues, mainly with *C. elegans.* The gene encoding the YEATS domain containing protein 4 is an orthologue of the *C. elegans* gene gfl-1, involved in processes such as nematode larval development, embryo development, RNA interference and negative regulation of vulval development. The gene encoding a protein of the Spec3 superfamily is an orthologue of the *C. elegans* gene Y97E10AL.1, related to male gonads. The gene encoding the phosphatidylinositol n-acetylglucosaminyltransferase subunit is an orthologue of human pigc (phosphatidylinositol glycan anchor biosynthesis class C), a component of the membrane. The gene encoding a protein with an Ost 4 superfamily motif (g14400) is an uncharacterized protein orthologue of the human ost4 gene (oligosaccharyltransferase complex subunit 4 (non-catalytic)). Finally, the gene encoding for the mitochondrial Cytochrome c oxidase subunit 4 isoform 1, was identified as a *Toxocara canis* orthologue (gene Tcan_14798). Unfortunately, none of the down regulated genes detected in the two hour induction had annotation or had clear orthologues with genes in the databases (NCBI, Wormbase or Uniport).

Selection in differentially expressed genes

For the last objective, we evaluated how selection is affecting genes related to the early stages of the host-pathogen interaction, modelled through the induction with insect tissues. We analyzed the patterns of selection in the differentially expressed genes. None of the three differentially expressed genes in the one hour induction had significant selection signatures (Directional or balancing selection). However, one of the two hour induction genes had a significant selection signature, negative D value of -1.1117. This gene, g12790, was the one encoding the Spec3 superfamily protein, related to male gonad.

We also characterized selection patterns in genes encoding differentially expressed proteins. These data belong to the proteomic part of the *S. carpocapsae* project currently ongoing in our group. The proteomic study correspond to a second induction of IJ nematodes with insect tissues (*G. mellonella* intestines or hemolymph) for four hours [Rougon-Cardoso *et al.,* 2016; Flores-Ponce *et al.,* submitted].

There were 1301 differentially expressed proteins, from which 561 were differential exclusively under induction with intestine, 550 only with hemolymph and 190 with both conditions. We additionally analyzed 2,237 expressed proteins that were not differential (with no difference in the expression between samples and control). The genes encoding for these proteins were then compared with the results obtained from the analyses of selection with both interspecific and intraspecific data. This not only allowed us to evaluate the patterns of selection but to see if these patterns were different in relation to the patterns in non-differentially expressed proteins, and in relation to the global scans.

We obtained interspecific data for 590 of the protein-coding genes, from which 143 were differentially expressed and 447 non-differentially expressed. From the 143 differentially expressed, 11 (7.69%) had sites under positive selection (see appendix 4). In the genes that were not differentially expressed, we found 30 (6.71%) with sites under positive selection (see appendix 6).

In the interspecific analysis we obtained evolutionary data for 3093 proteins (1009 differential and 2084 non-differential). Among the differentially expressed proteins, 42

(4.16%) presented significant Tajima D values in their coding genes. From these, 23 (2.28%) presented positive D values, and 19 (1.88%) presented negative D values. In non-differentially expressed proteins, we found 79 (3.80%) significant D values, 29 (1.4%) of them with positive values, and 50 (2.4%) with negative values. A summary of all these values including the ones obtained in the global scans are presented in Table 14.

Table 14. Patterns of selection in the genome of *S. carpocapsae* and in identified expressed proteins

Interspecific analysis	**N**	**Positive selection [a]**		
Global	1552	74 (4.77%)		
Differentially expressed	143	11 (7.69%)		
Non-differentially expressed	447	30 (6.71%)		
Intraspecific analysis	**N**	**Significant D [b]**	**Significant positive D [c]**	**Significant negative D [c]**
Global	14994	530 (3.53%)	228 (1.52%)	302 (2.01%)
Differentially expressed	1009	42 (4.16%)	23 (2.28%)	19 (1.88%)
Non-differentially expressed	2084	79 (3.80%)	29 (1.40%)	50 (2.40%)

N, number of genes analyzed.
a, Branch-sites test, LRT, $p<0.05$.
b, Tajima's D, 96% confidence level obtained from the real distribution.
c, Tajima's D, 98% confidence level obtained from the real distribution.

Discussion

This project had the main goal of exploring the impact that a parasitic lifestyle could have in nematode genomes, specifically in the entomopathogenic nematode *S. carpocapsae*. We had three particular goals for which we characterized selection patterns at a genome level in parasitic and free-living nematodes, searched for genes responding to the early stages of the interaction with host tissues, and finally looked to see if these genes showed specific patterns of selection.

Positive selection in interspecific data

For the interspecific analysis, we compared *S. carpocapsae*'s genome with a vertebrate parasite and closer free-living species of the phylogenetic clade IV. To assess the consistency of the observed patterns, we analyzed another set of nematodes with similar lifestyles, parasitic and free-living, but from a different phylogenetic clade (V).

This comparison between species had some limitations. First, only coding regions of the genome could be analyzed, and the genomes available at the time of the study was low and at different stages of completeness. Also, nematode species are in general highly divergent [Parkinson *et al.*, 2004] which makes it difficult to find a high number of one-to-one orthologues.

However, we obtained comparable data sets, even though they represent a low percentage of the genome, 7% and 9% of the protein-coding genes each. As low percentage of orthologous proteins were expected. In previous analysis, 245 orthologs were found in a set of nine nematode species including six of the species used in this project [Rougon-Cardoso *et al.*, 2016]. This could be improved with the analysis of more closely related species, leading to the identification of higher numbers of orthologues. A possible set of species to compare with is the recently sequenced genomes of Strongyloides and its related facultative parasitic specie [Hunt et al., 2016]. Also, genes analyzed at this level are among the most conserved, and it can be counterintuitive to look for positive selection in them. Still, they could be of special interest because

variation is shown at the sequence level but they preserve their function. This makes it relevant to evaluate if the observed differences between the species were fixed by positive selection or by random drift [Kimura 1968; King 1696]. As proposed by the neutral theory, the mayor expectance is the action of neutrality, however, it has been reported that changes in conserved genes are driven by positive selection more frequently than by neutrality [Bazykin and Kondrashov, 2012].

Even considering that we scanned a relatively low number of conserved genes, we could detect genes with sites evolving under positive selection, genes evolving under purifying selection, and a few genes evolving under neutrality. Proportions of genes evolving under purifying selection and neutrality seem to be constant in the branch model test. We identified many genes under purifying selection and a low proportion of genes evolving under neutrality. This is consistent with the reduction of variation by purifying selection acting against new mutations.

Vertebrate parasites showed some variation in the proportion of genes with sites evolving under positive selection when comparing between clades (Table 5). This variation could be due to their evolutionary history. In 2015, Blaxter and Koutsovoulos proposed a model in which one of the origins of vertebrate parasitism arose from entomopathogenic nematodes. This based on the phylogenetic associations between Strongylomorphs with Heterorhabditidae, and between Strongyloidoidae and Steinernematidae [Blaxter and Koutsovoulos, 2015]. However, a more recent study [Hunt *et al*., 2016] proposed the transition from free-living lifestyle through facultative parasitism to obligate parasitism in the *Strongyloides* genus.

In the case of the entomopathogenic or free-living nematodes, variation was not observed. It appears to be consistent with the convergent evolution in entomopathogens which has produced similar morphology and lifestyles, including their bacterial association with symbiotic bacteria and host interactions [Poinar, 1993; Blaxter and Koutsovoulos, 2015].

Yet, the observed proportions in genes with positively selected sites found in parasites and in free-living nematodes were similar in both clades. The free-living nematodes had a statistically significant higher proportion of genes with sites under positive selection (5.81% in free-living and 3.49% in parasitic nematodes); however, functional enrichment tests showed a higher number of enriched GO terms in parasites. This gives additional support to previous observations indicating that an arms-race interaction between hosts and pathogens affects specific genetic functions but not necessarily increases the global number of positively selected genes [Rougon-Cardoso *et al.*, 2016].

In parasites, most of the enriched functions involved immune response, antimicrobial production, and mitochondrial processes (Appendix 2 and 3). This last one is interesting because there are few examples in which genes related to mitochondrial function might be linked to parasitism [Rougon-Cardoso *et al.*, 2016]. One of these examples is the *C. elegans* homologue gene ATAD-3, which is suggested to be important for the increase in mitochondrial activity during the transition to later larval stages [Hoffmann *et al.,* 2009], and the defective mitochondrial respiration family member protein 1, which is involved in the regulation of growth rate. *S. carpocapsae* needs to go through developmental changes to establish itself in the insect body during the pathogenic process, which might explain the relevance of genes like this in nematode parasitism.

In this work, the mitochondrial enriched functions were associated to two genes, one in *S. carpocapsae* encoding for an ATP synthase B-like protein which is an orthologue of the *asb-1* gene in *C. elegans*. A*sb-1* is part of the mitochondrial respiratory chain complex V required for embryonic viability and normal proliferation of germline cells [Hu and Barr, 2005]. The other gene encodes for a RNA helicase domain containing protein in *Haemonchus contortus*. The *C. elegans* orthologue is predicted to have hydrolase activity and participates in embryo development and viability [Eki *et al.,* 2007]. Thus, regulation of embryo development and the timing of transition to later larval states might be relevant for nematode parasitism, a suggestion that needs to be further investigated.

Selection in intraspecific data

The intraspecific approach allowed us to conduct a wider analysis of the effect that lifestyle might have at genomic level. It included coding and intergenic regions and could detect more recent or ongoing selection events [Nielsen 2005]. This kind of analysis can be influenced by demographic processes, which might produce signals that can be confused with signatures of selection. However, these demographic events are expected to affect all genes in a genome, while selection pressures will act in specific loci of the genome [Tajima, 1989; Galtier *et al*., 2000; Stajich and Hahn, 2005]. It has previously been reported that better results can be achieved if specific population parameters are modeled for the populations under analysis [Pybus *et al*., 2015]. However, since nematode population variation has not been well documented we could not use that kind of approach.

Therefore, conducting a genome-wide level analysis will help to discern demographic effects from the action of selection over gene's variability. Our strategy consisted in estimating Tajima´s D in 1000 bp non-overlapping sliding windows covering the whole genome. It has been reported that the window size is not critical when detected selection signals [Lo *et al*., 2016]. However, windows must contain a sufficient number of polymorphisms for obtaining reasonable estimates of the neutrality test. Using large windows will dilute the selection signal due to the increase of number of neutral sites. The disadvantage of choosing a window too small is fail in detecting this signal if the selection signal is linked to a longer region [Lo *et al.,* 2016; Croze *et al*., 2017]. Finding the optimum window size could balance the noise reduction with selection signal identification to maximize power of the tests. Additionally, uniform window size is not appropriate when looking for genetic parameters such as recombination rate and linkage disequilibrium which vary across the genome [Beissinger *et al.,* 2015]. Also, we used non-overlapping (distinct) windows to separate the genome into fragments of equal length. Using non-overlapping windows helps reducing the sampling error reducing the number if statistical test performed. Some of the windows lacked coverage in some strains and were discarded from the analysis. With the data from the windows analyzed, we obtained an observed distribution. From this distribution, we estimated corrected

confidence limits to detect Tajima D values that were significantly different from both the expected values under neutrality and the values linked to demographic processes. Such demographic processes are assumed to be reflected in the average of D values across all the analyzed windows.

The comparison of these confidence limits with those obtained from the theoretical distribution proposed by Tajima [Tajima 1989] indicated that our approach avoided a bias in accepting or rejecting candidate regions. As seen in Figure 7 the Tajima confidence limits are asymmetrical in relation to the real data distribution. Proving that the estimation of confidence limits from the real distribution of D values improves the power of the test and avoids bias towards positive or negative values. This correction was implemented based on the manuscript by Schmidt and Pool in 2002, where they obtained a similar result using simulations for different demographic scenarios [Schmidt and Pool 2002]. One problem of our approach is that when doing comparisons between species, the percentage of significant positive and negative values will always be similar because the idea is to accept a predefined marginal percentage of values from each side of the distribution (in our case, 2%). This nullifies any possible conclusion regarding the impact of lifestyle in the proportion of genes with signatures of selection. However, the comparison of the distribution itself can gives us some clues regarding the impact of lifestyle in the patterns of selection at the genomic level. The shape of the distribution of *S. carpocapsae* is different and more skewed to positive values of Tajima's D in relation to the distribution of *C. briggsae* (Figure 9). This might be indicative of pervasive balancing selection acting in the genome of the entomopathogenic nematode, due to trench warfare coevolutionary host-pathogen dynamics. One characteristic of this dynamic is the promotion of balancing selection [Tellier *et al*., 2014]. Until now, the examples of genes showing signatures of these evolutionary genetic interactions were limited to single genes, mostly related to effectors or genes related to the immune response. An example is the gene *msp1* encoding for the merozoite surface protein 1 (Msp1) in the malaria parasite *Plasmodium falciparum* [Conway *et al*., 2000; Aguileta *et al*., 2009]. This gene has evolved under balancing selection, showing highly divergent alleles with stable frequencies in endemic populations [Conway *et al*., 2000].

One alternative explanation for our results is that demographic processes are not only affecting the average D value but also changing the shape of the distribution. This is a more complex scenario that will need to be further investigated. One possibility to study this is to compare nematode populations with specific demographic histories, such as population reduction or expansion, or bottlenecks. Another way could be using *S. carpocapsae* data to compute simulations of demographic processes.

Another problem in our comparison is that the genetic distance among the *C. briggsae* strains is bigger than the distance among the *S. carpocapsae* strains included in the analyses (Appendix 7). This is reflected in the significantly higher number of invariable windows detected in *S. carpocapsae* in relation to *C. briggsae* (χ^2 test, $p< 6.1x10^{-72}$). However, positive values of Tajima's D correlate with higher levels of nucleotide diversity in relation to segregating sites [Tajima 1989]. Therefore, the fact that *S. carpocapsae* shows more windows with positive D values than *C. briggsae* when there is less genetic distance among their strains, reinforces the idea that balancing selection is acting with more strength in *S. carpocapsae* due to its pathogenic lifestyle. Under balancing selection, more alleles will be maintained in each segregating site. To maintain different alleles in a significant number of these sites, a higher strength of selection is required.

Differentially expressed genes due to the interaction with insect tissues

Identifying the time when IJs reach the insect's intestines and hemocoel was the key to the infection model proposed during this study. After tracking the presence of nematodes at different times post-infection, we found that 3 hours post infection is the first time point where IJs can be found inside the insect larvae. Even though we found minimal nematode presence after 4 hours post infection, most of the IJs were found in the intestine (Figure 11). After this time we found nematodes in both tissues, with a prevalence in the larvae intestine. The presence of nematodes in the hemocoel only after 4 hours post infection, could suggest that *S. carpocapsae*'s entrance to *G. mellonella* is mainly through mouth an anus. If the IJs entrance occurs equally through all of the host's natural openings, we would have expected to see a higher number of

nematodes in the hemocoel at 3-4 hours post infection. A way to resolve the exact time of arrival to either the intestine or hemocoel could be repeating the experiment with a much higher number of infected larvae per time point. This could be challenging because obtaining large number of larvae in the same stage (same size and weight) requires several *G. mellonella* cultures, which was not possible to maintain in our lab.

Nevertheless, the results from the infection experiment let us propose a possible model for the early stages of infection (Figure 12). In this model, the time where IJs reach the intestines is 3 hours after the first contact with the larvae. Then the proposed time for IJs to reach the hemocoel is an hour later (4 hours post infection). Therefore, a one hour induction with larval intestine could simulate the arrival and transition to hemocoel. The same way, induction with larvae hemolymph would simulate the arrival to hemocoel and the interaction with the insect immune system.

During the inductions, we tried to minimize the possible extra variation that could affect the RNA extraction. These included interaction with *X. nematophila* or other bacteria present in the nematode cuticles, which were eliminated with the washes prior to the induction and with the use of nalidixic acid during the incubation with insect tissue. The original induction protocol used ampicillin, which showed low efficiency in killing *X. nematophila* when performing an antibiogram. The most effective antibiotic inhibiting *X. nematophila*'s growth was nalidixic acid.

Also, we tried to reduce the variation by doing the replicates under the same conditions on consecutive days. What we did not account for is that this disposition caused a batch effect, which could be in part why RNA replicate samples had a big variation among them, limiting the detection of differential expressed genes. When adding this effect in the experimental design we did no saw power increased in the differential expression analysis, as reported in the edgR manual. The experiment needs to be planed differently taking looking for more sources of variation that could had interfered in the experiment.

This only allowed us to identify three genes with different expressions at the earliest time point (one hour induction). Interestingly the only gene that was up-regulated, a c-type

lectin domain-containing protein 160, with a *C. elegans* orthologue that at the same time is an orthologue of the human gene *FCER2* [Spencer *et al.,* 2011]. This gene encodes a protein that is a B-cell specific antigen and a low-affinity receptor for IgE. When binding to a high-affinity IgE-Fc receptor (Fc3RI) stimulates immune responses important for defense against bacterial, viral, and parasitic infections [Wurzburg *et al*., 2006]. Being associated with immune response activity in humans, it can suggest that it could be involved in regulating immune reactivity in *S. carpocapsae*.

The other two differentially expressed genes at one hour after induction had no obvious participation in the infection process. One had a tyrosine kinase activity with a possible participation in cell proliferation [Schlessinger, 2000]; and the other encodes an aldehyde dehydrogenase with a role in aldehyde detoxification in the cell [Crabb *et al*., 2004]. The rest of the differentially expresses genes (at two hour after induction) had no obvious participation in the infection process either. These genes encoded for a broad type of proteins. One involved in larval development, one related to male gonad; another a membrane component; one part of oligosaccharyltransferase complex; and the mitochondrial Cytochrome c oxidase subunit 4 isoform 1. A possibility is that these genes could be participating in the resuming of *S. carpocapsae* functions and life cycle.

To expand our assessment of how selection is affecting genes related to the host-pathogen interaction we analyzed differentially expressed proteins in early stages of the interaction with the insect host. Proteomic data came from an induction experiment with a longer incubation time, four hours with hemolymph, with two replicates [Rougon-Cardoso *et al.,* 2016; Flores-Ponce *et al.,* submitted]. Replicates in this experiment were consistent among them, attributable possibly to a longer exposure to the hemolymph.

The number of differentially expressed proteins identified, correspond to a 7.96% of the estimated genes for *S. carpocapsae*. Since analyzed proteins correspond only to the soluble fraction, the real number of proteins expressed when interacting with the insect hemolymph could be higher. Further analysis of the insoluble fraction should be planned to optimize the results of this proteomic approach.

Selection in differentially expressed genes

To evaluate if the genes and proteins participating in the interaction with the host show different patterns of selection, we compared the proportion of genes with signatures of selection in differentially expressed genes and proteins against the proportion found in the global analyses and with the proportion in non-differentially expressed coding-proteins genes. Table 14 resumes the results for the differential expressed proteins. Since the number of differentially expressed genes was low, due to high variation among replicates, those results are only described below.

From the transcriptomic approach, only one gene of the 13 identified (from both time points) had a significant signature of selection. This one gene represents a higher proportion of proteins with patterns of selection (7.69%) in relation to the global analysis (3.53%), although it is not significantly different from the proportions observed in proteins in the interspecific and intraspecific analysis (Fisher's exact test, p=0.473208, and p=0.374067, respectively).

Then, we proceeded to do the comparison with the genes encoding the differentially expressed proteins. In the interspecific analysis, the proportion of genes with positive selected sites was 1.2 to 1.6 times higher in differentially expressed proteins (7.69%) in relation to both, the global analysis (4.77%) and the non-differentially expressed proteins (6.71%), although the differences were not statistically significant (Fisher´s exact test, p=0.15624, and p=0.706381, respectively). In this analysis, we detected ancient selection [Nielsen 2005] and it is possible that the genes that were targeted at that time are different than the genes that are currently relevant in the interaction with the host. Yet, a tendency for an increased proportion of genes with signatures of positive selection in differentially expressed proteins seems to remain. A possible way to investigate this is to analyze a less phylogenetical distant set of species, this to cover different temporal windows of *S. carpocapsae*'s evolutionary history.

In the intraspecific analyses, we observed a slightly higher proportion of genes with significant D values among the differentially expressed genes (4.16%) in relation to the global scan (3.53%) or to the non-differentially expressed genes (3.80%). Although these differences were also not significant (Fisher's exact test, p=0.292928, and

p=0.621698, respectively). It is interesting that the differences were more evident in the proportion of genes with significant positive D values. There were 2.28% of these genes among the differentially expressed genes, 1.52% in the global scan, and 1.40% in the non-differentially expressed genes (an increase of 1.46 and 1.67 times, respectively). The differences were not significant but showed marginal p-values (Fisher's exact test, p=0.066236, and p=0.075006, respectively). It is also interesting that the tendency is reverted in the proportion of genes with significant negative D values, which is lower in differentially expressed genes (1.88%) than in the global scan (2.01%) or in the non-differentially expressed genes (2.40%). The tendency for an increased proportion of genes with positive D values in differentially expressed genes cannot be explained by demographic processes because these processes would have affected the non-differentially expressed genes in the same manner as the differentially expressed ones. Additionally, a higher proportion of balancing selection and a lower proportion of directional selection in differentially expressed genes indicates that balancing selections is more relevant than other types of selection in the current genes mediating host-pathogen interactions.

These results imply that the dynamic polymorphism models might be dominant and more general than the arms-race model. Also, since there were no enriched functions found under these model, our results could suggest that this coevolutionary dynamic does not target specific genetic functions in contrast to the arms-race model. A few of the genes reported to be evolving under balancing selection in parasites are directly related to host interaction, such as avoiding host recognition, or production of toxins [Aguileta *et al*., 2009].

Here, besides the gene encoding an immunoglobulin i-set domain containing protein, we did not find proteins with obvious participation in the interaction with the host. However, we found some genes associated with mitochondrial localization or biological processes. Two of them are under balancing selection, a mitochondrial membrane carrier homolog, and a gene encoding for a subunit of the mitochondrial NADH dehydrogenase (ubiquinone) complex (complex I), which had one of the most positive D values. Other genes under balancing selection were associated with embryo

development and transport biological processes. Thus, a more detailed analysis of the genes with signatures of balancing selection might be useful to identify novel relevant factors in the pathogenic process.

Conclusions

- This project provides a first portrait of the effects that lifestyle might have in shaping the patterns of selection at the genomic level in the entomopathogenic nematode *S. carpocapsae*.
- The performed genome-wide scans for positive selection indicates that in pathogenic nematodes, positive selection is targeting specific genetic functions, possibly due to an arms-race host-pathogen interaction.
- The intraspecific genomic analysis indicates that balancing selection could be acting with more strength in the *S. carpocapsae* genome in comparison with the genome of free-living *C. briggsae*.
- Since the Tajima's D was corrected to reduce demographic effects, the increased effect of balancing selection could be attributable to a trench warfare coevolutionary host-pathogen dynamic.
- The intraspecific genome-wide scans indicate that in *S. carpocapsae*, trench warfare dynamic through balancing selection is not targeting specific genetic functions.
- Conserved genes involved in the growth and development regulation, and mitochondrial function were targets of selection in the past, possibly due to host-pathogen dynamics (arms-race)
- Differentially expressed proteins responding to the interaction with host tissues are slightly enriched for balancing selection.

References

Aguileta G, Refregier G, Yockteng R, Fournier E, Giraud T. Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. Infect Genet Evol 2009;9:656-670.

Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. Available online at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Balasubramanian N, Hao YJ, Toubarro D, Nascimento G, Simões N. Purification, biochemical and molecular analysis of a chymotrypsin protease with prophenoloxidase suppression activity from the entomopathogenic nematode *Steinernema carpocaps*ae. Int J Parasitol. 2009;39(9):975-84.

Bazykin GA, Kondrashov AS. Major role of positive selection in the evolution of conservative segments of Drosophila proteins. Proc Biol Sci. 2012;279(1742):3409-17.

Bedding R. Low cost in vitro mass production of *Neoaplectana* and *Heterorhabditis* species (Nematoda) for field control of insect pests. Nematologica. 1981;27:109-114.

Beissinger TM, Rosa GJ, Kaeppler SM, Gianola D, de Leon N. Defining window-boundaries for genomic analyses using smoothing spline techniques. Genet Sel Evol. 2015;47:30.

Blaxter M, Koutsovoulos G. The evolution of parasitism in Nematoda. Parasitology. 2015;142:S26-S39.

Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, Vierstraete A, et al. A molecular evolutionary framework for the phylum Nematoda. Nature. 1998;392: 71–75.

Blaxter ML. Nematoda: genes, genomes and the evolution of parasitism. Adv Parasitol. 2003;54, 101–195.

Boemare N, Laumond C, Mauleon H. The entomopathogenic nematode-bacterium complex: biology, life cycle and vertebrate safety. Biocontrol SciTechn. 1996;6(3): 333-346

Boemare N. Biology, Taxonomy, and Systematics of *Photorabdus* and *Xenorhabdus*. in Gaugler I, ed. Entomopathogenic Nematology. CABI Publishing. New Jersey. 2002 p 57-78.

Bromham L, Cowman PF, Lanfear R. Parasitic plants have increased rates of molecular evolution across all three genomes. BMC Evol Biol. 2013;13:126.

Burman, M. *Neoaplectana carpocapsae*: Toxin production by axenic insect parasitic nematodes. Nematologica. 1982;28:62-70.

Burnell AM, Stock P. *Heterorhabditis*, *Steinernema* and their bacterial symbionts, lethal pathogens of insects. Nematology. 2000; 2:31-42.

Castillo JC, Reynolds SE, Eleftherianos I. Insect immune responses to nematode parasites. Trends Parasitol. 2011;27(12):537-47.

Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540-552.

Clausen CP. Phoresy Among Entomophagous Insects. Annu Rev Entomology. 1976;21:1-414.

Coghlan A. Nematode Genome Evolution. In: WormBook (ed. The *C. elegans* Research Community) 2005. http://www.wormbook.org.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21(18):3674–3676.

Conway DJ, Fanello C, Lloyd JM, Al-Joubori BM, Baloch AH, Somanath SD, Roper C, Oduola AM, Mulder B, Povoa MM, Singh B, Thomas AW. Origin of *Plasmodium falciparum* malaria is traced by mitochondrial DNA. Mol Biochem Parasitol. 2000;111(1):163-71.

Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. Mol Cell Proteomics. 2014;13:2513–2526.

Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26:1367–1372.

Crabb DW, Matsumoto M, Chang D, You M. Overview of the role of alcohol dehydrogenase and aldehyde dehydrogenase and their variants in the genesis of alcohol-related pathology. Proc Nutr Soc. 2004;63(1):49-63.

Croze M, Wollstein A, Božičević V, Živković D, Stephan W, Hutter S. A genome-wide scan for genes under balancing selection in Drosophila melanogaster. BMC Evol Biol. 2017; 17: 15.

Cutter AD. Molecular evolution inferences from the *C. elegans* genome. WormBook, ed. The *C. elegans* Research Community, WormBook, (March 5 2010) doi/10.1895/wormbook.1.7.1, http://www.wormbook.org.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, Depristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–2158.

Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 2012;9(8):772.

Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011;27:1164-1165.

Davis EL, Hussey RS, Baum TJ, Bakker J, Schots A, Rosso MN, Abad P. Nematode parasitism genes. Annu Rev Phytopathol. 2000;38: 365 -396.

Davis EL, Hussey RS, Baum TJ. Getting to the roots of parasitism by nematodes. Trends Parasitol. 2004;20:134–141

Dillman AR, Macchietto M, Porter CF, Rogers A, Williams B, Antoshechkin I, Lee MM, Goodwin Z, Lu X, Lewis EE, Goodrich-Blair H, Stock SP, Adams BJ, Sternberg PW, Mortazavi A. Comparative genomics of *Steinernema* reveals deeply conserved gene regulatory networks. Genome Biology. 2015;16:1-21.

Dobon A, Bunting DC, Cabrera-Quio LE, Uauy C, Saunders DG. The host-pathogen interaction between wheat and yellow rust induces temporally coordinated waves of gene expression. BMC Genomics. 2016;17(1):380.

Dorris M, De Ley P, Blaxter ML. Molecular Analysis of Nematode Diversity and the Evolution of Parasitism. Pararitol Today. 1999;15(5):188-193.

Duret L. Neutral theory: The null hypothesis of molecular evolution. Nat Education. 2008;1(1):218.

Ehlers RU. Mass production of entomopathogenic nematodes for plant protection. Appl Microbiol Biotechnol. 2001;56:623–633.

Eki T, Ishihara T, Katsura I, Hanaoka F. A Genome-wide Survey and Systematic RNAi-based Characterization of Helicase-like Genes in *Caenorhabditis elegans*. DNA Res. 2007;14(4):183-199.

Ellegren H. Comparative genomics and the study of evolution by natural selection. Mol Ecol. 2008;17:4586–4596

Fu W, Akey JM. Selection and Adaptation in the Human Genome. Annu Rev Genom Hum G. 2013;14:467–89.

Galtier N, Depaulis F, Barton NH. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics. 2000;155(2):981-987.

Gaugler R. Entomopathogenic nematology. CABI Publishing. New Jersey. 2002.

Giblin-Davis RM, Kanzaki N, Davies KA. Nematodes that Ride Insects: Unforeseen consequences of Arriving Species. Fla Entomol. 2013;96(3):770-780.

Gotz P, Boman A, Boman HG. 1981. Interactions between Insect Immunity and an Insect-Pathogenic Nematode with Symbiotic Bacteria. Proceedings of the Royal Society of London. Series B. Biological Sciences 212:333-350.

Griffin CT, Boemare NE, Lewis EE. Biology and Behaviour in Grewal PS, Ehlers RU, Shapiro-Ilan DI, eds. Nematodes as Biocontrol Agents. CABI Publishing. 2005 p 47-64.

Guindon S, Gascuel O. A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. Syst Biol. 2003;52:696-704.

Hamilton MB. Population genetics. West Sussex: Wiley-Blackwell; 2009.

Hao Y, Montiel R, Nascimento G, Toubarro D, Simões N. Identification, characterization of functional candidate genes for host–parasite interactions in entomopathogenetic nematode *Steinernema carpocapsae* by suppressive subtractive hybridization. Parasitol Res. 2008;103:671-683.

Hao YJ, Montiel R, Abubucker S, Mitreva M, Simões N. Transcripts analysis of the entomopathogenic nematode *Steinernema carpocapsae* induced in vitro with insect haemolymph. Mol Biochem Parasitol. 2010;169:79-86.

Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, De La Cruz N, Davis P, Duesbury M, Fang R, Fernandes J, Han M, Kishore R, Lee R, Müller HM, Nakamura C, Ozersky P, Petcherski A, Rangarajan A, Rogers A, Schindelman G, Schwarz EM, Tuli MA, Van Auken K, Wang D, Wang X, Williams G, Yook K, Durbin R, Stein LD, Spieth J, Sternberg PW: WormBase: a comprehensive resource for nematode research. Nucleic Acids Res. 2010;38:D463-D467.

Hoffmann M, Bellance N, Rossignol R, Koopman WJK, Willems PHGM, Mayatepek E, Bossinger O, Distelmaier F. *C. elegans* ATAD-3 is essential for mitochondrial activity and development. PLoS ONE. 2009;4:

Holterman M, van der Wurff A, van den Elsen S, van Megen H, Bongers T, Holovachov O, Bakker J, Helder J. Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades. Mol Biol Evol. 2006;23:1792-1800.

Hu J, Barr MM. ATP-2 interacts with the PLAT domain of LOV-1 and is involved in *Caenorhabditis elegans* polycystin signaling. Mol Biol Cell. 2005;16:458-69. doi:10.1091/mbc.E04-09-0851

Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. Genetics. 1987;116(1):153-9.

Hunt VL, Tsai IJ, Coghlan A, Reid AJ, Holroyd N, Foth BJ, Tracey A, Cotton JA, Stanley EJ, Beasley H, Bennett HM, Brooks K, Harsha B, Kajitani R, Kulkarni A, Harbecke D, Nagayasu E, Nichol S, Ogura Y, Quail MA, Randle N, Xia D, Brattig NW, Soblik H, Ribeiro DM, Sanchez-Flores A, Hayashi T, Itoh T, Denver DR, Grant W, Stoltzfus JD, Lok JB, Murayama H, Wastling J, Streit A, Kikuchi T, Viney M, Berriman M. The genomic basis of parasitism in the *Strongyloides* clade of nematodes. Nat Genet. 2016;48(3):299–307.

Jackson DE. Chemical coevolution: host-parasite arms race runs hot and cold. Curr Biol. 2008;18(7):R306-8.

Jones JT, Moens M, Mota M, Li H, Kikuchi T. *Bursaphelenchus xylophilus*: opportunities in comparative genomics and molecular host–parasite interactions. Mol Plant Pathol. 2008;9:357–368.

Kaya HK, Gaugler R. Entomopathogenic nematodes. Annu Rev Entomology. 1993;38:181-206.

Kaya HK, Stock SP. Techniques in insect nematology. In: Manual of Techniques in Insect Pathology, L. A. Lacey, editors. Academic Press, London; 1997. p. 281-324.

Kimura M. Evolutionary rate at the molecular level. Nature. 1968;217:624-626.

King JL, Jukes TH. Non-Darwinian Evolution. Science. 1969;164:788-97.

Kitano T, Sumiyama K, Shiroishi T, Saitou N. Conserved evolution of the Rh50 gene compared to its homologous Rh blood group gene. Biochem Biophys Res Commun. 1998;249(1):78-85.

Kitano T, Saitou N. Evolution of Rh blood group genes have experienced gene conversions and positive selection. J Mol Evol. 1999;49(5):615-26.

Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9:357-359.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23:2947-2948.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009;25:1754-60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N., Marth G., Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009; 25:2078-9.

Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-93.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;arXiv:1303.3997 [http://arxiv.org/abs/1303.3997]

Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178-2189.

Li X, Cowles EA, Cowles RS, Gaugler R, Cox-Foster DL. Characterization of immunosuppressive surface coat proteins from Steinernema glaseri that selectively kill blood cells in susceptible hosts. Mol Biochem Parasit.. 2009;165:162–169.

Lo CL, Lossie AC, Liang T, Liu Y, Xuei X, Lumeng L, Zhou FC, Muir WM. High Resolution Genomic Scans Reveal Genetic Architecture Controlling Alcohol Preference in Bidirectionally Selected Rat Model. PLoS Genet. 2016;12(8):e1006178.

Mbata GN, Shapiro-Ilan DI. Compatibility of *Heterorhabditis indica* (Rhabditida: Heterorhabditidae) and *Habrobracon hebeto*r (Hymenoptera: Braconidae) for biological control of *Plodia interpunctella* (Lepidoptera: Pyralidae). Biol Control. 2010;54:75-82.

McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40:4288-4297.

McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature. 1991;351(6328):652-4.

Nei M, Miller JC. A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. Genetics. 1990;125(4):873-879.

Nei M, Suzuki Y, Nozawa M. The Neutral Theory of Molecular Evolution in the Genomic Era. Annu Rev Genom Hum G. 2010;11:265–89.

Nei M. Selectionism and Neutralism in Molecular Evolution. Mol Biol Evol. 2005;22(12):2318–2342.

New World Encyclopedia contributors, "Symbiosis," New World Encyclopedia, http://www.newworldencyclopedia.org/p/index.php?title=Symbiosis&oldid=945948 (accessed January, 2017).

Nielsen R. Molecular signatures of natural selection. Annu Rev Genet. 2005;39:197-218.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, J Sninsky J, Adams MD, Cargill M. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 2005 Jun;3(6):e170.

Parkinson J, Mitreva M, Whitton C, Thomson M, Daub J, Martin J, Schmid R, Hall N, Barrell B, Waterston RH, McCarter JP, Blaxter ML. A transcriptomic analysis of the phylum Nematoda. Nat Genet. 2004;36(12):1259-1267.

Platt HM. Foreword. In: Lorenzen, S. ed. The Phylogenetic Systematics of Free-living Nematodes. 1994;pp. i–ii, .

Poinar GO. Nematodes for Biological Control of Insects. Boca Raton, Florida: CRC Press. 1979.

Poinar GO. Origins and phylogenetic relationships of the entomophilic rhabditis, *Heterorhabditis* and *Steinernema*. Fundam Appl Nematol. 1993;16(4): 333-338.

Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J, Engelken J. Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. Bioinformatics. 2015;31(24):3946–3952.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org (2015).

Ridley M. Evolution. 3rd Edition. Blackwell Publishing. 2003.

Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Meth. 2013; 10:71-73.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139-140.

Rougon-Cardoso DA, Flores-Ponce M, Ramos-Aboites HE, Martinez-Guerrero CE, Hao Y-J, Cunha L, Rodríguez-Martinez JA, Ovando-Vázquez C, Bermúdez-Barrientos JR, Abreu-Goodger C, Chavarria-Hernández N, Simões N, Montiel R. The genome, transcriptome, and proteome of the nematode *Steinernema carpocapsae*: evolutionary signatures of a pathogenic lifestyle. Sci Rep. 2016;6.

Sambrook J, Fritsch EF, Maniatis T. Molecular cloning: a laboratory manual. 2nd edition. New York: Cold Spring Harbor Laboratory Press; 1989.

Sawyer SL, Emerman M, Malik HS. Ancient Adaptive Evolution of the Primate Antiviral DNA-Editing Enzyme APOBEC3G. PLoS Biol. 2004 Sep; 2(9): e275

Schlessinger J. Cell Signaling by Receptor Tyrosine Kinases. Cell. 2000;103(2):211 – 225.

Schmidt D, Pool J. The effect of population history on the distribution of the Tajima's D statistic. New York: Cornell University Press. 2002. http://wolfweb.unr.edu/~drschmidt/TajimasD.pdf

Schmidt HA, Strimmer K, Vingron M, von Haeseler A.TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 2002;18:502-504.

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. Nature. 2011;473:337–342.

Shapiro-Ilan DI, Gaugler R. Production technology for entomopathogenic nematodes and their bacterial symbionts. J Ind Microbiol Biot. 2002;28:137-146.

Simões N, Caldas C, Rosa JS, Bonifassi E, Laumond C. Pathogenicity caused by high virulent and low virulent strains of *Steinernema carpocapsae* to *Galleria mellonella*. J Invertebr Pathol 2000;75:47-54.

Simões N, Caldas C, Rosa JS, Bonifassi E, Laumond C. Pathogenicity Caused by High Virulent and Low Virulent Strains of *Steinernema carpocapsae* to *Galleria mellonella*. J Invertebr Pathol. 2000;75:74-54.

Simões N, Rosa JS. Pathogenicity and Host Specificity of Entomopathogenic Nematodes. Biocontrol Sci Techn. 1996;6(3):403- 412.

Smart GC. Entomopathogenic nematodes for the biological control of insects. J Nematol. 1995;27:529-534.

Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL, McWhirter RD, Petersen SC, Sreedharan VT, Widmer C., Jo J, Reinke V, Petrella L, Strome S, Von Stetina S, Katz M, Shaham S, Raetsch G, Miller DM. A spatial and temporal map of *C. elegans* gene expression. Genome Res. 2011; 21:325-41.

Stahl EA, Bishop JG. Plant-pathogen arms races at the molecular level. Curr Opin Plant Biol. 2000 Aug;3(4):299-304.

Stajich JE, Hahn MW. Disentangling the Effects of Demography and Selection in Human History. Mol Biol Evol. 2005;2(1):63-73.

Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, Coulson A, D'Eustachio P, Fitch DH, Fulton LA, Fulton RE, Griffiths-Jones S, Harris TW, Hillier LW, Kamath R, Kuwabara PE, Mardis ER, Marra MA, Miner TL, Minx P, Mullikin JC, Plumb RW, Rogers J, Schein JE, Sohrmann M, Spieth J, Stajich JE, Wei C, Willey D, Wilson RK, Durbin R,

Waterston RH.The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. PLoS Biol. 2003;1(2):166-192.

Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123:585–595.

Tanaka T, Nei M. Positive darwinian selection observed at the variable-region genes of immunoglobulins. Mol. Biol. Evol. 1989;6:447–459.

Tellier A, Moreno-Gámez S, Stephan W. Speed of adaptation and genomic footprints of host-parasite coevolution under arms race and trench warfare dynamics. Evolution. 2014;68(8):2211-24.

Thomas CG, Wang W, Jovelin R, Ghosh R, Lomasko T, Trinh Q, Kruglyak L, Stein LD, Cutter AD. Full-genome evolutionary histories of selfing, splitting, and selection in *Caenorhabditis*. Genome Res. 2015;25:667-678.

Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673-4680.

Toubarro D, Avila MM, Hao Y, Balasubramanian N, Jing Y, Montiel R, Faria TQ, Brito RM, Simões N. A Serpin Released by an Entomopathogen Impairs Clot Formation in Insect Defense System. PLoS ONE. 2013;8(7): e69161.

Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016;13(9):731-40.

Wernersson R, Pedersen AG. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 2003;31:3537-3539.

Woolhouse ME, Webster JP, Domingo E, Charlesworth B, Levin BR.. Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nat Genet. 2002;32(4):569-77.

Wu W, Goodman M, Lomax MI, Grossman LI. Molecular evolution of cytochrome c oxidase subunit IV: evidence for positive selection in simian primates. J Mol Evol. 1997;44(5):477-91.

Wurzburg BA, Tarchevskaya SS, Jardetzky TS. Structural Changes in the Lectin Domain of CD23, the Low-Affinity IgE Receptor, upon Calcium Binding. Structure. 2006;14:1049–1058.

Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. TREE. 2000;15:496-503.

Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 2002;19:908–917.

Yang Z, Nielsen R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 1998;46:409-418.

Yang Z. Likelihood ratio test for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol.1998;15(5)568-573.

Yang Z. Maximum Likelihood Estimation on Large Phylogenies and Analysis of Adaptive Evolution in Human Influenza Virus A. J Mol Evol. 2000;51:423–432.

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586-1591.

Yarwood CE. Obligate Parasitism. Annu Rev Plant Physiol. 1956;7:115-142.

Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 2005;22:2472-2479.

Appendix 1. Commands and parameters

Clarisse and the Codeml parameters

Clarisse

```
clarisse -v –t8 --config config --template plantilla {1..9}*
```

Template

```
          noisy = 3

        verbose = 0
        runmode = 0

        seqtype = 1
      CodonFreq = 2
          clock = 0
         aaDist = 0

       *NSsites = 0

          icode = 0

      fix_alpha = 1
          alpha = 0
         Malpha = 0
          ncatG = 10

          getSE = 0
  RateAncestor = 0

  Small_Diff = .5e-
                6
     cleandata = 1
         method = 1
```

Config files

Branch

```
1:
 seqfile: =*.phy
 treefile: ../tree.nwk
 kappa: 0
 fix_kappa: 0
 omega: 0
 fix_omega: 0
 model: 0
 outfile: 0.out
2:
 seqfile: =*.phy
 treefile: ../c4n4_tree.nwk
 fix_kappa: 1
 kappa: =1|kappa
 fix_omega: 0
 omega: 0
 model: 0
 outfile: 1.out
3:
 seqfile: =*.phy
 treefile: ../c4n4_tree_1.nwk
 fix_kappa: 1
 kappa: =1|kappa
 fix_omega: 0
 omega: 0
 model: 2
 outfile: 2.out
4:
 seqfile: =*.phy
 treefile: ../c4n4_tree_1.nwk
 fix_kappa: 1
 kappa: =1|kappa
 fix_omega: 1
 omega: 1
 model: 2
 outfile: 3.out
5:
 seqfile: =*.phy
 treefile: ../c4n4_tree.nwk
 fix_kappa: 1
 kappa: =1|kappa
 fix_omega: 0
 omega: 0
 model: 1
 outfile: 4.out
```

Branch-Site

```
1:
 seqfile: =*.phy
 treefile: ../c4n4_tree_1.nwk
 fix_kappa: 0
 kappa: 0
 fix_omega: 0
 omega: 0
 model: 2
 NSsites: 2
 outfile: 1_bs.out
2:
 seqfile: =*.phy
 treefile: ../ tree_1.nwk
 fix_kappa: 0
 kappa: 0
 fix_omega: 1
 omega: 1
 model: 2
 NSsites: 2
 outfile: 2_bs.out
```

Index built

```
bowtie2-build reference_file.fasta bt2_index_base
```

Mapping and SNP determination

Index reference:

```
bwa index -a bwtsw reference.fasta
```

-a Burrows–Wheeler transform construction algorithm: bwtsw or is [auto]

Mapping with mem algorithm:

```
bwa mem -M -A 2 -a -t 8 reference.fasta FileR1.fastq FileR2.fastq > Alnpe_Az20.sam
```

-M mark shorter split hits as secondary
-A score for a sequence match, which scales options
-a output all alignments for SE or unpaired PE
-t number of threads

Filtered with awk

```
awk '$3!="*"' Alnpe_Az20.sam > awk_Alnpe_Az20.sam
```

View:

```
samtools view -bS awk_Alnpe_Az20.sam > awk_Alnpe_Az20.bam
```

-b change the output format from the default SAM to BAM.
-S Specify a single input file format option.

Sort:

```
samtools sort awk_Alnpe_Az20.bam -o awk_Alnpe_Az20.sorted.bam
```

-o Write final output to FILE rather than standard output

Index:

```
samtools index awk_Alnpe_Az20.sorted.bam
```

Remove duplicates:

```
samtools rmdup -S awk_Alnpe_Az20.sorted.bam Az20.bam
```

-s Remove duplicates for single-end reads
-S Treat paired-end reads and single-end reads.

mpileup:

```
samtools mpileup -ug -q 20 -f reference.fasta Az154.bam > Az154.bcf
```

-u generate uncompressed VCF/BCF output.
-g generate genotype likelihoods in BCF format.
-q skip alignments with mapQ smaller than 20.
-f faidx indexed reference sequence file.

View:

```
bcftools view -vcg Az154.bcf > Az154.vcf
```

-v select/exclude comma-separated list of variant types: snps,indels,mnps,other [null]
-c minimum/maximum count for non-reference (nref), 1st alternate (alt1), least frequent (minor), most frequent (major) or sum of all but most frequent (nonmajor) alleles [nref]
-g require one or more hom/het/missing genotype or, if prefixed with "^", exclude sites with hom/het/missing genotypes

Call:

```
bcftools call Az154.bcf -m -O v -o Az154_call2.vcf
```

-m alternative model for multiallelic and rare-variant calling.
-O output type: 'v' uncompressed VCF.
-o write output to a file.

Quality analysis

```
fastqc  -o FastQC  --extract  -f fastq  -t 4  File.fastq
```

-o Creates all output files in the specified output directory.
-f sequence file format detection.
-t Specifies the number of files which can be processed simultaneously.
--extract Zipped output file will be uncompressed in the same directory after it has been created.

Mapping with bowtie

```
bowtie2 -p 4 --al -x transcritosSC --un-conc-gz  file.unaln.gz -1 File1.fastq
-2 File2.fastq |  samtools view -Sb - > File.bam
```

-p Launch a specified number of parallel search threads
--al Write unpaired reads that align at least once to a file.

-x the base name of the index for the reference genome/transcriptome
-1 file or files containing mate 1 of paired-end reads
-2 files or files containif mate 2s of paired-end reads
--un-conc-gz Write paired-end reads that fail to align concordantly to file(s)

Samtools view
-S Ignored for compatibility with previous samtools versions
-b Output in the BAM format

Quantification of transcript abundance

```
express target_seqs.fasta aligned_reads.(sam/bam)  -o output_dir
```

-o Sets the name of the directory in which eXpress will write all of its output.

Appendix 2. Enriched GO terms in the genes with sites under positive selection in nematode clade IV

Organism	Shared with	GO-ID	Term	Category	FDR	P-Value
Sc	Sr, Pr	GO:0005198	structural molecule activity	F	1.14E-02	3.08E-05
Sc		GO:0000975	regulatory region DNA binding	F	8.72E-01	4.73E-02
Sc		GO:0002920	regulation of humoral immune response	P	8.72E-01	4.73E-02
Sc		GO:0002922	positive regulation of humoral immune response	P	8.72E-01	4.73E-02
Sc		GO:0002697	regulation of immune effector process	P	8.72E-01	4.73E-02
Sc		GO:0002699	positive regulation of immune effector process	P	8.72E-01	4.73E-02
Sc		GO:0002225	positive regulation of antimicrobial peptide production	P	8.72E-01	4.73E-02
Sc		GO:0002831	regulation of response to biotic stimulus	P	8.72E-01	4.73E-02
Sc		GO:0002833	positive regulation of response to biotic stimulus	P	8.72E-01	4.73E-02
Sc		GO:0001067	regulatory region nucleic acid binding	F	8.72E-01	4.73E-02
Sc		GO:0002700	regulation of production of molecular mediator of immune response	P	8.72E-01	4.73E-02
Sc		GO:0002702	positive regulation of production of molecular mediator of immune	P	8.72E-01	4.73E-02
Sc		GO:0002760	positive regulation of antimicrobial humoral response	P	8.72E-01	4.73E-02
Sc		GO:0002759	regulation of antimicrobial humoral response	P	8.72E-01	4.73E-02
Sc		GO:0002440	production of molecular mediator of immune response	P	8.72E-01	4.73E-02
Sc		GO:0002684	positive regulation of immune system process	P	8.72E-01	4.73E-02
Sc		GO:0000276	mitochondrial proton-transporting ATP synthase complex, coupling	C	8.72E-01	4.73E-02
Sc		GO:0002784	regulation of antimicrobial peptide production	P	8.72E-01	4.73E-02
Sc		GO:0002775	antimicrobial peptide production	P	8.72E-01	4.73E-02
Sc		GO:0000808	origin recognition complex	C	8.72E-01	4.73E-02
Sr	Sc, Pr	GO:0005198	structural molecule activity	F	3.44E-01	1.03E-03
Sr		GO:0005275	amine transmembrane transporter activity	F	1.00E+00	1.73E-02
Pr	Sc, Sr	GO:0005198	structural molecule activity	F	4.08E-03	1.18E-05
Pr		GO:0000976	transcription regulatory region sequence-specific DNA binding	F	5.97E-01	3.46E-03
Pr		GO:0005244	voltage-gated ion channel activity	F	9.68E-01	9.89E-03
Pr		GO:0000977	RNA polymerase II regulatory region sequence-specific DNA binding	F	9.68E-01	1.90E-02
Pr		GO:0001012	RNA polymerase II regulatory region DNA binding	F	9.68E-01	1.90E-02
Pr		GO:0003676	nucleic acid binding	F	9.68E-01	4.63E-02

Pr= *P. redivivus*, Sc = *S. carpocapsae*, and Sr = *S. ratti*. Number of genes analyzed: Sc = 74, Sr= 24 , and Pr= 91 .Category: F= Moluecular function, P= Biological process, and C= Cellular component

Appendix 3. Enriched GO terms in the genes with sites under positive selection in nematode clade V

Organism	Shared	GO-ID	Term	Category	FDR	P-Value
Hb		GO:0005198	structural molecule activity	F	1.31E-01	4.96E-04
Hb	Hc	GO:0001948	glycoprotein binding	F	1.00E+00	4.59E-02
Hb	Hc	GO:0005604	basement membrane	C	1.00E+00	4.59E-02
Hb	Hc	GO:0005605	basal lamina	C	1.00E+00	4.59E-02
Hb		GO:0000054	ribosomal subunit export from nucleus	P	1.00E+00	4.59E-02
Hb		GO:0000291	nuclear-transcribed mRNA catabolic process, exonucleolytic	P	1.00E+00	4.59E-02
Hb	Hc	GO:0002162	dystroglycan binding	F	1.00E+00	4.59E-02
Hb		GO:0001727	lipid kinase activity	F	1.00E+00	4.59E-02
Hb		GO:0000149	SNARE binding	F	1.00E+00	4.59E-02
Hb		GO:0001101	response to acid chemical	P	1.00E+00	4.59E-02
Hc		GO:0005737	cytoplasm	C	5.96E-01	7.81E-03
Hc		GO:0000278	mitotic cell cycle	P	5.96E-01	2.41E-02
Hc		GO:0005681	spliceosomal complex	C	5.96E-01	2.41E-02
Hc		GO:0001664	G-protein coupled receptor binding	F	5.96E-01	2.41E-02
Hc	Hb	GO:0001948	glycoprotein binding	F	5.96E-01	3.63E-02
Hc		GO:0001711	endodermal cell fate commitment	P	5.96E-01	3.63E-02
Hc		GO:0001714	endodermal cell fate specification	P	5.96E-01	3.63E-02
Hc		GO:0000959	mitochondrial RNA metabolic process	P	5.96E-01	3.63E-02
Hc		GO:0000957	mitochondrial RNA catabolic process	P	5.96E-01	3.63E-02
Hc	Hb	GO:0005604	basement membrane	C	5.96E-01	3.63E-02
Hc	Hb	GO:0005605	basal lamina	C	5.96E-01	3.63E-02
Hc		GO:0000963	mitochondrial RNA processing	P	5.96E-01	3.63E-02
Hc		GO:0000960	regulation of mitochondrial RNA catabolic process	P	5.96E-01	3.63E-02
Hc		GO:0000177	cytoplasmic exosome (RNase complex)	C	5.96E-01	3.63E-02
Hc	Hb	GO:0002162	dystroglycan binding	F	5.96E-01	3.63E-02
Hc		GO:0002039	p53 binding	F	5.96E-01	3.63E-02
Hc		GO:0004683	calmodulin-dependent protein kinase activity	F	5.96E-01	3.63E-02
Hc		GO:0003724	RNA helicase activity	F	5.96E-01	3.63E-02
Hc		GO:0000132	establishment of mitotic spindle orientation	P	5.96E-01	3.63E-02
Hc		GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	P	5.96E-01	3.95E-02
Hc		GO:0000375	RNA splicing, via transesterification reactions	P	5.96E-01	3.95E-02
Hc		GO:0003676	nucleic acid binding	F	6.22E-01	4.31E-02

Hb = *H. bacteriophora*, Hc = *H. contortus*, and Cb = *C. briggsae*. Number of genes analyzed: Hb = 61, Hc = 55, and Cb = 87. Category: F= Moluecular function, P= Biological process, and C= Cellular component.

Appendix 4. Genes for differentially expressed proteins with sites under positive selection

Differentially expressed

Name	Seq. Description	GOs
g1350.t1	triosephosphate isomerase	P:cellular component assembly involved in morphogenesis; P:cellular homeostasis; P:multicellular organismal aging; C:cell part; F:triose-phosphate isomerase activity; P:fructose metabolic process; P:mannose metabolic process; P:inositol metabolic process; P:gluconeogenesis; P:glycolysis; P:carbon utilization; P:glycerolipid metabolic process
g1898.t1	succinate dehydrogenase cytochrome b560 subunit	C:mitochondrial respiratory chain; P:respiratory electron transport chain; P:regulation of growth rate; P:endocytosis; F:iron ion binding; P:embryo development; F:succinate dehydrogenase activity; C:succinate dehydrogenase complex; F:electron carrier activity; P:electron transport; P:oxidative phosphorylation; P:benzoate metabolic process; P:reductive tricarboxylic acid cycle
g4565.t1	sodium-independent organic anion transporter family protein	C:cell part; P:embryo development; F:protein binding; F:transporter activity; P:transport; C:membrane
g5030.t1	thymosin beta-4	F:actin binding; C:cytoplasm; P:cytoskeleton organization
g5568.t1	protein lin-32	P:transcription, DNA-dependent; C:intracellular membrane-bounded organelle; F:protein dimerization activity
g6502.t1	sorting nexin-13	C:cell part; P:termination of G-protein coupled receptor signaling pathway; F:protein binding; F:phosphatidylinositol binding
g11410.t1	60s ribosomal protein l4	P:embryo development; P:multicellular organismal aging; F:binding; F:structural constituent of ribosome; C:ribosome; P:translation; P:ribosome biogenesis
g11641.t1	tata-box binding protein	P:multi-organism process; P:larval development; C:transcription factor complex; F:regulatory region DNA binding; P:multicellular organismal aging; F:nucleic acid binding transcription factor activity; C:pronucleus; F:sequence-specific DNA binding; P:male gamete generation; P:regulation of growth rate; P:endocytosis; F:transcription factor binding; P:transcription from RNA polymerase II promoter; P:embryo development
g12399.t1	malate dehydrogenase	P:malate metabolic process; P:cellular carbohydrate metabolic process; F:L-malate dehydrogenase activity; P:pyruvate metabolic process; P:reductive tricarboxylic acid cycle; P:glyoxylate metabolic process
g12781.t1	serine threonine kinase	P:anterior/posterior axis specification; P:establishment or maintenance of cell polarity; P:asymmetric protein localization; P:vulval development; C:cytoplasmic part; P:genitalia development; P:embryo development; F:ATP binding; P:protein phosphorylation; F:protein serine/threonine kinase activity; F:protein tyrosine kinase activity; P:serine family amino acid metabolic process
g15020.t1	nuclear receptor-binding protein	F:ATP binding; P:protein phosphorylation; F:serine-type endopeptidase inhibitor activity; F:protein tyrosine kinase activity

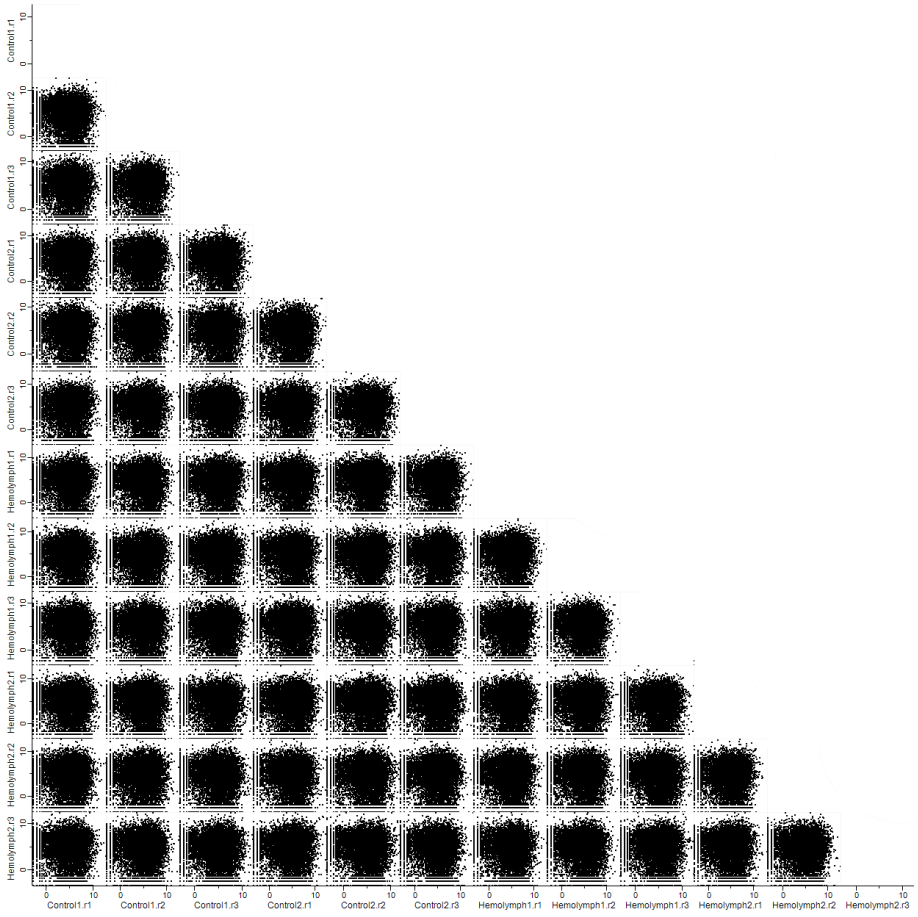
Positive selection, Branch-sites test, LRT, $p<0.05$

Appendix 5. Correlation analysis of RNA-Seq samples

Correlation among samples

	C 1.r1	C 1.r2	C 1.r3	C 2.r1	C 2.r2	C 2.r3	H1.r1	H1.r2	H1.r3	H2.r1	H2.r2	H2.r3
C 1.r1	1.000											
C 1.r2	0.006	1.000										
C 1.r3	0.003	0.002	1.000									
C 2.r1	0.008	-0.017	0.006	1.000								
C 2.r2	0.003	0.002	0.011	0.050	1.000							
C 2.r3	0.032	0.001	-0.015	0.010	0.023	1.000						
H1.r1	-0.005	0.003	-0.001	0.003	-0.004	0.032	1.000					
H1.r2	-0.016	0.012	0.025	0.001	0.011	-0.017	0.019	1.000				
H1.r3	-0.015	0.004	-0.005	0.023	-0.004	0.002	0.017	-0.001	1.000			
H2.r1	-0.003	-0.005	-0.002	0.013	-0.006	0.014	0.015	0.002	-0.015	1.000		
H2.r2	-0.012	0.007	-0.012	-0.007	-0.005	-0.025	0.000	0.001	-0.008	0.029	1.000	
H2.r3	-0.004	0.003	0.003	0.077	0.020	0.095	0.000	-0.004	0.037	0.005	0.002	1.000

C = Control, H =Hemolymph



Multi scatter plot in log scale for the replicates.

Appendix 6. Genes for non-differentially expressed proteins with sites under positive selection

Non-differentially

Name	Seq. Description	GOs
g701.t1	protein tnt- isoform b	P:actomyosin structure organization; P:somatic muscle development; P:female gamete generation; P:axonogenesis; P:locomotion; P:endocytosis; P:muscle contraction; C:troponin complex
g1112.t1	copine family protein	F:protein binding
g2246.t1	60s ribosomal protein l22	P:embryo development; P:regulation of growth rate; P:multicellular organism growth; P:larval development; F:structural constituent of ribosome; C:ribosome; P:translation; P:ribosome biogenesis
g2284.t1	small subunit ribosomal protein 14	F:structural constituent of ribosome; C:ribosome; P:translation; P:ribosome biogenesis
g2865.t1	mitochondrial dicarboxylate carrier	C:organelle inner membrane; C:intrinsic to membrane; P:transport
g3137.t1	26s protease regulatory subunit 4	F:nucleoside-triphosphatase activity; F:peptidase activity, acting on L-amino acid peptides; P:cellular response to stress; C:proteasome regulatory particle; P:multicellular organismal aging; P:mitotic spindle organization; C:ribonucleoprotein complex; P:gene expression; F:structural molecule activity; P:embryo development; F:ATP binding; C:cytoplasm; P:protein catabolic process
g3601.t1	protein isoform d	C:intrinsic to membrane; P:transport; F:hydrolase activity
g4088.t1	hypothetical tyrosinase-like protein in chromosome	F:binding; P:metabolic process; F:oxidoreductase activity
g4593.t1	NAD mitochondrial	P:hydrogen transport; C:intracellular membrane-bounded organelle; F:NAD(P)+ transhydrogenase (AB-specific) activity; C:integral to membrane; F:NADP binding; P:oxidation-reduction process; P:nicotinamide metabolic process; P:nicotinate nucleotide metabolic process
g5114.t1	methionyl-tRNA synthetase mitochondrial	P:embryo development; F:ATP binding; F:methionine-tRNA ligase activity; C:cytoplasm; P:methionyl-tRNA aminoacylation; P:methionine metabolic process
g5888.t1	small ubiquitin-related modifier precursor	P:protein-based cuticle development; P:gastrulation; P:synaptonemal complex organization; P:protein localization to organelle; P:vulval development; P:regulation of growth rate; P:genitalia development; P:transcription from RNA polymerase II promoter; F:protein binding
g6166.t1	proteasome subunit alpha type-5	P:embryo development; P:multicellular organismal aging; C:proteasome core complex; C:intracellular membrane-bounded organelle; P:larval development; P:ubiquitin-dependent protein catabolic process; F:threonine-type endopeptidase activity
g6716.t1	upf0160 protein mitochondrial-like	0
g6931.t1	protein isoform a	F:lipid binding
g7631.t1	NADH-ubiquinone oxidoreductase b22 subunit	P:endocytosis; P:regulation of growth rate; P:larval development
g9282.t1	60s ribosomal protein l24	P:embryo development; P:regulation of growth rate; P:gene expression; C:ribonucleoprotein complex; P:larval development; F:structural molecule activity

Name	Seq. Description	GOs
g9593.t1	large subunit ribosomal protein 32	F:structural constituent of ribosome; C:ribosome; P:translation; P:ribosome biogenesis
g9630.t1	mitochondrial aspartate transaminase	F:transaminase activity; P:cellular amino acid metabolic process; P:biosynthetic process; F:pyridoxal phosphate binding
g9736.t1	paramyosin	P:primary metabolic process; P:multicellular organismal reproductive behavior; C:myosin complex; F:hydrolase activity, acting on glycosyl bonds; P:larval development; C:contractile fiber; P:regulation of organelle organization; F:protein binding; P:endocytosis; P:skeletal myofibril assembly; F:motor activity
g10232.t1	60s acidic ribosomal protein p0	F:binding; F:carbon-oxygen lyase activity; P:ribosome biogenesis; F:structural constituent of ribosome; C:ribosome; P:translational elongation
g11412.t1	60s ribosomal protein l17	P:larval development; P:axonogenesis; P:regulation of growth rate; P:embryo development; F:structural constituent of ribosome; P:translation; C:large ribosomal subunit; P:ribosome biogenesis
g11772.t1	golgi reassembly-stacking protein 2	F:protein binding
g12180.t1	atp synthase b-like protein	P:male mating behavior; F:hydrogen ion transmembrane transporter activity; P:multicellular organismal aging; P:regulation of growth rate; C:mitochondrial proton-transporting ATP synthase complex; P:embryo development; C:cell projection; P:vulval development; P:ATP synthesis coupled proton transport
g13097.t1	acyl- -binding protein	P:lipid localization; F:fatty-acyl-CoA binding
g13354.t1	40s ribosomal protein s4	F:RNA binding; F:structural constituent of ribosome; C:ribosome; P:translation; P:ribosome biogenesis
g13483.t1	cathepsin z precursor	P:embryo development; C:extracellular region part; F:endopeptidase activity; P:anatomical structure morphogenesis; P:molting cycle, protein-based cuticle; C:intrinsic to membrane; P:larval development; P:proteolysis; F:cysteine-type peptidase activity
g13933.t1	ribosomal protein s13	P:translation; P:viral genome expression; F:RNA binding; C:small ribosomal subunit; C:nuclear lumen; P:larval development; P:pancreas development; P:RNA splicing; P:embryo development; F:structural constituent of ribosome; P:ribosome biogenesis
g14568.t1	polyadenylate-binding protein 1-like isoform 1	P:embryo development; P:regulation of growth rate; C:ribonucleoprotein granule; P:genitalia development; P:tissue morphogenesis; F:RNA binding; P:sexual reproduction; P:endocytosis; F:nucleotide binding
g15076.t1	60s ribosomal protein l7	P:embryo development; P:regulation of growth rate; P:protein metabolic process; F:binding; P:gene expression; C:ribonucleoprotein complex; P:larval development; F:structural molecule activity
g15891.t1	60s ribosomal protein l18a	F:structural constituent of ribosome; C:ribosome; P:translation; P:ribosome biogenesis

Appendix 7. Distance Matrices

Matrices of Maximum Likelihood distances between the analysed strains of *S. carpocapsae* and *C. briggsae*.

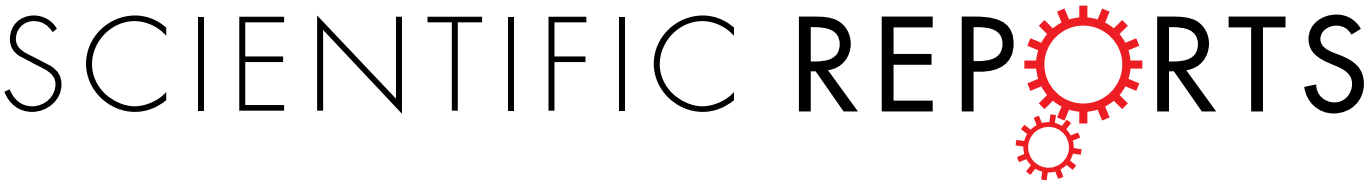
S. carpocapsae strains (best-fit model: HKY+G)

	ScPB	ScAll	ScAz20	ScAz154	ScAz157
ScPB	0				
ScAll	0.00013	0			
ScAz20	0.00015	0.00021	0		
ScAz154	0.00015	0.00021	0.00001	0	
ScAz157	0.00016	0.00021	0.00001	0.00001	0

C. briggsae strains (best-fit model: GTR+G)

	AF16	ED3101	JU1348	QR25	VX0034
AF16	0				
ED3101	0.24891	0			
JU1348	0.03257	0.27339	0		
QR25	0.19026	0.09148	0.19811	0	
VX0034	0.18476	0.4807	0.21348	0.29113	0

Best-fit model estimated with jModelTest v.2.1.3 [Darriba *et al.,* 2012]; ML distances estimated with Tree-Puzzle v.5.3.rc16 [Schmidt *et al*., 2002]



OPEN

Received: 17 May 2016
Accepted: 31 October 2016
Published: 23 November 2016

The genome, transcriptome, and proteome of the nematode *Steinernema carpocapsae*: evolutionary signatures of a pathogenic lifestyle

Alejandra Rougon-Cardoso[1,2,*], Mitzi Flores-Ponce[1,*], Hilda Eréndira Ramos-Aboites[1], Christian Eduardo Martínez-Guerrero[1], You-Jin Hao[3], Luis Cunha[4], Jonathan Alejandro Rodríguez-Martínez[2], Cesaré Ovando-Vázquez[1], José Roberto Bermúdez-Barrientos[1], Cei Abreu-Goodger[1], Norberto Chavarría-Hernández[5], Nelson Simões[6] & Rafael Montiel[1]

The entomopathogenic nematode *Steinernema carpocapsae* has been widely used for the biological control of insect pests. It shares a symbiotic relationship with the bacterium *Xenorhabdus nematophila*, and is emerging as a genetic model to study symbiosis and pathogenesis. We obtained a high-quality draft of the nematode's genome comprising 84,613,633 bp in 347 scaffolds, with an N50 of 1.24 Mb. To improve annotation, we sequenced both short and long RNA and conducted shotgun proteomic analyses. *S. carpocapsae* shares orthologous genes with other parasitic nematodes that are absent in the free-living nematode *C. elegans*, it has ncRNA families that are enriched in parasites, and expresses proteins putatively associated with parasitism and pathogenesis, suggesting an active role for the nematode during the pathogenic process. Host and parasites might engage in a co-evolutionary arms-race dynamic with genes participating in their interaction showing signatures of positive selection. Our analyses indicate that the consequence of this arms race is better characterized by positive selection altering specific functions instead of just increasing the number of positively selected genes, adding a new perspective to these co-evolutionary theories. We identified a protein, ATAD-3, that suggests a relevant role for mitochondrial function in the evolution and mechanisms of nematode parasitism.

Global losses due to pests can vary from about 26 to 80% depending on the type of crop[1]. Chemical pesticides are commonly used to fight this problem, however, they pose threats to humans, wildlife, and might have an adverse impact on soil fertility by killing beneficial microorganisms[2]. Other strategies rely on biological control agents, but their use is not generalized because of their limited efficiency when compared to pesticides. Genetic improvements are possible, especially when genomic information of the biological agent is available[3,4]. Entomopathogenic nematodes (EPNs) from the family of *Steinernematidae* have been commercialized in many countries as a

[1]Laboratorio Nacional de Genómica para la Biodiversidad, Unidad de Genómica Avanzada, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional. Km 9.6 Libramiento Norte Carretera Irapuato-León, C.P. 36821 Irapuato, Guanajuato, Mexico. [2]Laboratory of Agrogenomic Sciences, Universidad Nacional Autónoma de México (UNAM), ENES-León, 37684, León, Guanajuato, Mexico. [3]College of Life Science, ChongQing Normal University, ChongQing 401331, China. [4]Cardiff School of Biosciences, Cardiff University, Park Place, Sir Martin Evans Building, Museum Avenue, Cardiff, Wales CF10 3US, UK. [5]Cuerpo Académico de Biotecnología Agroalimentaria. Instituto de Ciencias Agropecuarias, Universidad Autónoma del Estado de Hidalgo. Av. Universidad Km 1, Rancho Universitario, Tulancingo de Bravo, Hidalgo, C.P. 43600, Mexico. [6]CIRN/Departamento de Biologia, Universidade dos Açores, Rua Mãe de Deus, 13. 9500-321 Ponta Delgada. S. Miguel-Açores, Portugal. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to R.M. (email: rafael.montiel@cinvestav.mx)

Strain	Source	Library type	Platform	No. of runs	No. of reads (Millions)	Average read length (bp)	Insert size (bp)
Breton	DNA	Shotgun	454 GS FLX	3	33.41	357	—
		Paired-end	454 GS FLX	2	27.85	334	8000
		Shotgun	SOLiD 5500xl	~Half lane	24.94	75	—
	RNA	Shotgun cDNA	454 GS FLX +	Half plate	0.09	288	—
		Paired-end cDNA	Illumina MiSeq	1	15.18	201	—
	sRNA	Shotgun	Illumina HiSeq 2500	1 lane (6 tagged libraries)	42.81	51	—
All	DNA	Paired-end	Illumina Genome Analyzer IIx	1	85.76	75	400
		Paired-end	Illumina Genome Analyzer IIx	1	103.09	100	350
		Paired-end	Illumina HiSeq 2000	1	131.59	100	1800
	RNA	Paired-end cDNA	Illumina Genome Analyzer IIx	4 (each from a different developmental stage)	260.86	75	200

Table 1. Summary of sequencing data from *Steinernema carpocapsae* strain Breton, compared to the sequencing data of the strain All[11].

biological insecticide for agricultural and horticultural crops and have attracted considerable attention because they are also potential models for symbiosis and pathogenesis[5]. One of the most well-known is *Steinernema carpocapsae* that shares a symbiotic relationship with the bacterium *Xenorhabdus nematophila*. Since it was thought that the bacteria were the main contributor to insect death, most research has focused on the pathogenic effect of the bacteria rather than the nematode[6–8]. Nevertheless, growing evidence suggests a more active role of the nematode in the pathogenic process[9,10]. In fact, a set of expanded gene families that are likely involved in parasitism were predicted in a recent genome analysis of *Steinernema* species[11]. Further genomic characterization will help to better understand the evolution and the function of these genomes in the symbiotic and pathogenic contexts. Parasitism is a common way of life among nematodes that has independently arisen at least 15 times during their evolution[12]. Particularly interesting are the phylogenetic associations between non-vertebrate and vertebrate parasites. The entomopathogenic Steinernematidae are phylogenetically related to Strongyloidoidae (Tylenchina; Panagrolaimomorpha), which infect mammals, suggesting a transition to vertebrate parasitism through host shifting[12]. The study of parasitism in *S. carpocapsae* should help to understand the origin and mechanisms of Strongyloidoids parasites, with implications for human health. For this study we produced a high-quality draft of the genome of *S. carpocapsae* strain Breton, and compared it with a recently published genome from a different strain of this species[11]. We further assessed the genetic signatures of its adaptation to a pathogenic lifestyle, and characterized the transcriptome by RNA-Seq, including both messenger RNA (mRNA), and small RNA (sRNA). We also present the most complete characterization to date of the proteome, generated by shotgun proteomics, two-dimensional gel electrophoresis (2DE) and SDS-PAGE. Additionally we conducted genome-wide scans for signatures of natural selection. We found several distinctive features related to pathogenesis through a comparison with both pathogenic and free-living nematodes.

Results and Discussion

Genome sequencing. Total DNA was extracted from isolated nuclei from a near isogenic line (~96% of estimated homozygosity) of *Steinernema carpocapsae* strain Breton. The use of isolated nuclei reduces the amount of symbiont and mitochondrial DNA, and the isogenic line was generated to avoid the acknowledged problems posed by heterozygosity for accurate genome assembly[12]. From one 454 shotgun library sequenced in three 454 FLX runs, we obtained 3,340,915 total reads with an average length of 357 bp. From one 454 paired-end library, with an insert size of 8 Kb, sequenced in two 454 FLX runs, we obtained 2,784,713 total reads with an average read length of 334 bp at each fragment end. From a SOLiD shotgun library sequenced in half a lane of SOLiD 5500xl, we obtained 24,942,584 reads of 75 bp (Table 1). By combining these long, paired-end, and short reads, we obtained a coverage of 32-fold, considering a genome size of ~110 Mb estimated by both flow cytometry and genome assembly. The final draft consists of 84,613,633 base pairs in 347 scaffolds, with an N50 of 1.24 Mega bases and with the largest scaffold of 8.7 Mb. This represents a notable improvement over a recently published genome that is more fragmented, with a much lower N50 (~0.3 Mb) and with the largest scaffold of only 1.7 Mb (Table 2). The average GC-content was of 45.67%, with 6.99% of repetitive sequences (Supplementary Table S1).

We assessed the completeness of the genome by analysing 248 ultra-conserved core eukaryotic genes[13], obtaining 99.6% completeness considering partial genes and 99.2% for complete genes. These parameters indicated that our draft genome is of high quality, which gives us confidence in the genome annotation described below.

Genome annotation. From the repetitive elements, we identified 1,702 distinct retrotransposon sequences representing at least eight families. Four were long interspersed element (LINE) groups, Cr1 being the most abundant, and 588 were short interspersed elements (SINEs), of which 432 belong to the tRNA-RTE family. We identified only two long terminal repeats (LTRs): *Gypsy* and *Pao*. We also identified eight families of DNA transposons, comprising 1,202 sequences, of which *hAT-Ac* was the most abundant with 388 elements, followed by *TcMar-Tc1*, *Merlin*, and the rolling-circle *Helitron* (327, 106, and 105 elements, respectively).

Strain	Breton	All
Sequencing depth	32X	330X
Estimated genome size in megabases (GSAssembler 2.7)	111.3	85.6
(Flow Cytometry)	~110	Not determined
Number of scaffolds	347	1,578
Total number of base pairs within assembled scaffolds	84,613,633	86,127,942*
N50 Scaffold length (bp)	1,245,171	299,566
Largest scaffold (bp)	8,793,593	1,722,607
GC content of whole genome (%)	45.67	45.53
Repetitive sequences (%)	6.99	7.46
Proportion of genome that is coding (exonic) (%)	19.72	38.8*
Proportion of genome that is transcribed (exons + introns) (%)	42.09	50.31*
Number of putative coding genes	16,333	28,313
Number of non-coding RNAs	1,317	Not determined
Mean gene size (bp)	2,681	2,030
Mean coding sequence length per gene (bp)	1,257	1,046*
Average exon number per gene	6	5
Average gene exon length (bp)	222.37	212
Average gene intron length (bp)	145.44	194
GC content in coding regions (%)	52.49	51.86
Functionally annotated genes (according to BLAST2GO default parameters)	10,395 (63.6%)	Not determined

Table 2. Summary statistics of assembly and annotation of the genome of *Steinernema carpocapsae* strain Breton, compared to the assembly of the strain All[11]. *Calculated from version PRJNA202318.WBPS6 obtained from www.wormbase.org.

We collected RNA from pooled nematodes taken from all life cycle stages and subjected to various conditions (growing in larvae of two different insect species and on two different *in vitro* media, as described in Materials and Methods) in order to maximize the inclusion of condition-specific genes. We obtained 15,180,085 reads with an average length of 201 bp from an Illumina paired-end library on a MiSeq, and 92,231 reads with an average length of 288 bp from a 454 library on a partial 454 FLX + plate. After quality filtering, 94.93% of the reads mapped to the masked genome, suggesting a good reliability of the genome assembly. We performed genome-guided *de novo* assembly of the transcriptome that resulted in 21,457,711 bp of assembled transcripts (without introns). In order to identify protein-coding genes in the assembled genome, we assigned specific weights to different types of evidence to generate consensus gene calls (see Material and Methods). The current genome sequence and annotation is available at www.genomevolution.org (ID 33774), and at the NCBI GenBank (BioProject ID# 39853).

We identified 16,333 protein-coding genes with an average length of 1,257 bp, an average exon length of 222.37 bp, and an average of six exons per gene. We also identified 6,708 alternative transcripts and 5,725 truncated genes (defined as predicted protein-coding genes missing a start codon). We verified the protein expression of 3,773 predicted genes through mass spectrometry analysis (see below and Supplementary Table S2). The total number of predicted genes in this study is much lower than the previously predicted number of genes (28,313) for the strain "All" of the same species[11]. In the previous study, the heterozygosity was not reduced through the generation of an isogenic line, potentially negatively impacting on their genome assembly[12]. In addition, they performed gene prediction using Augustus[14] with parameters optimized only for *Caenorhabditis elegans*. However, these species diverged ~280 million years ago[15], making it difficult to accurately predict genes in *Steinernema* by solely using *C. elegans* gene models. To overcome this bias, we combined predictions using Hidden Markov Models (HMM) trained on *S. carpocapsae* gene structures with *ab initio* predictions, along with HMM homology-based predictions using *C. elegans* genes and *Brugia malayi* gene predictions (all the predictions were obtained with Augustus[14]). Although *B. malayi* has the same estimated time of divergence from *S. capocapsae* as *C. elegans*[15], it is a parasite and therefore might share some homologous genes with *S. carpocapsae* that are not represented in *C. elegans*. However, the full strength of our approach is given by combining predictions using gene models from these species with *ab initio* predictions. When we used the same annotation strategy as in Dillman *et al.*[11] using *C. elegans* models, we only obtained 14,188 protein coding genes. This is lower than the previous study, and also lower to the one we obtained with our combined strategy, in which the number of predicted genes probably increased due to the inclusion of *B. malayi* models, along with the *ab initio* predictions.

In summary, the use of an isogenic line (see Material and Methods), the combination of different sequencing platforms (Table 1), and an improved annotation strategy, resulted in a higher quality genome compared to a recent publication (Table 2). In any case, the studies used different strains of the nematode, and their goals were different, with Dillman *et al.*[10] focusing on comparing their genome to that of other species of *Steinernema*. In our study, we obtained a higher quality genome, included analyses of the proteome and small RNAs, and performed a genome-wide scan of positive selection.

The most abundant GO terms in predicted genes are shown in Fig. 1. Our analysis revealed 135 enriched GO terms in *S. carpocapsae* when compared to those in *C. elegans* (Fisher's exact test, FDR $\leq$ 0.05) (Supplementary Table S3). Many of these GO terms are involved in degradation, protein modification, binding and transport, and could

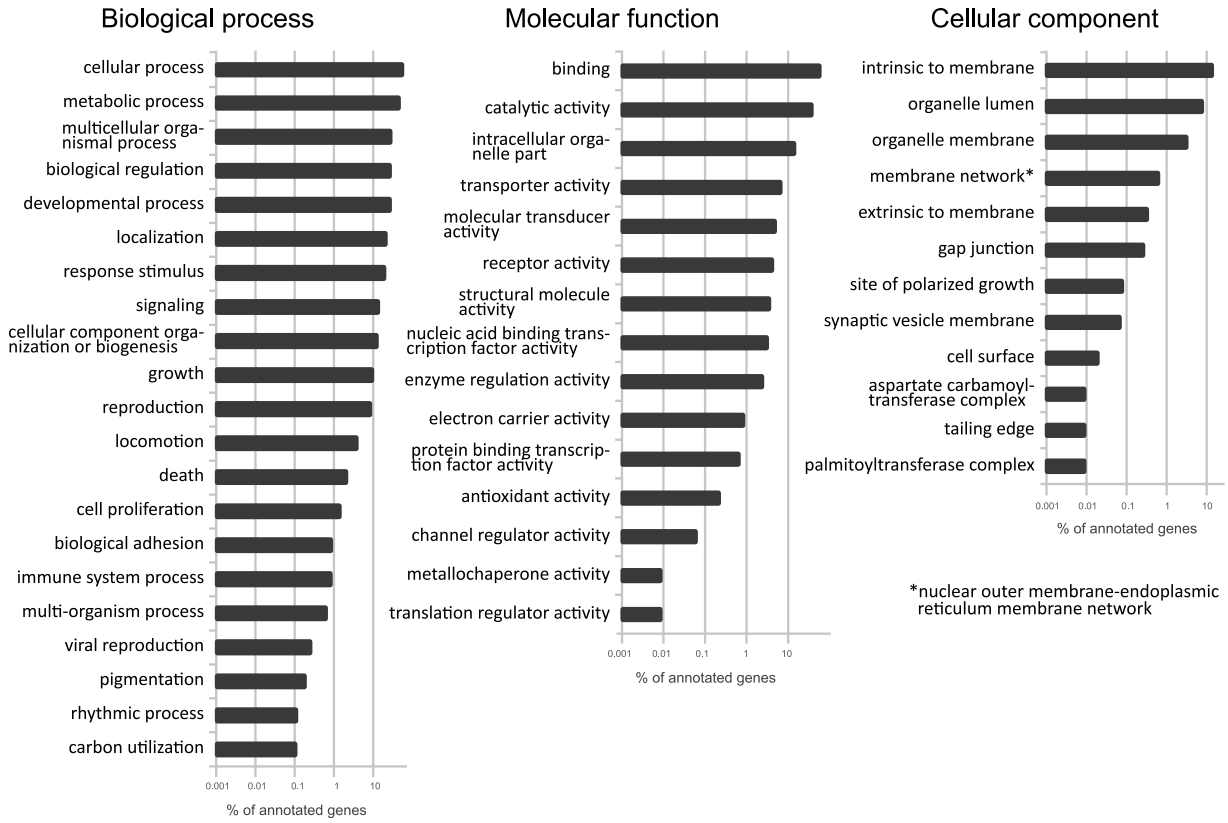


Figure 1. Enrichment analysis of GO terms in annotated sequences of *Steinernema carpocapsae,* in relation to those in *Caenorhabditis elegans*.

be associated to parasitism (reviewed in ref. 16). At least 22 GO terms are also enriched in at least two other pathogenic worms, but not in the free-living nematode *Pristionchus pacificus* (Table 3). Supplementary Table S4 shows the abundance of the different protein families in 10 different nematode genomes compared with the 30 most abundant families in *S. carpocapsae* (Supplementary Fig. S1). Most of the expanded families (16 out of 20) identified previously in the "All" strain of *S. carpocapsae*[11] are also overrepresented in our strain (Breton). In addition, Peptidase S1 is also overrepresented in *S. carpocapsae* and other parasites (*Bursaphelenchus xylophilus, Meloidogyne hapla,* and *M. incognita*) compared to *C. elegans*. Integrase, enoyl-acyl-carrier-protein reductase (ENR) (IPR014358), retrotransposon pao, and pimelyl-acyl-carrier protein methyl ester esterase PFAM domains, are also overrepresented in some parasites, including *S. carpocapsae*.

Since *C. elegans* and *S. carpocapsae* are phylogenetically distant from one another, we found no evident macro-synteny. However, there are genes located in single chromosomes of *C. elegans* that match genes located in single scaffolds of *S. carpocapsae.* A similar result was obtained in a comparison with the *Brugia malayi* genome (version WS253) (Fig. 2 and Table 4), even though this genome is not of the same quality as that of *C. elegans*. This reinforces the idea that the use of *B. malayi* gene models in the annotation strategy is at least as good as the use of *C. elegans* models.

Beyond the protein-coding potential of the genome, we predicted non-coding RNA (ncRNA) using a variety of tools (see Materials and Methods), identifying 1,097 tRNAs, 40 rRNAs (15 5S rRNA, 1 5.8S rRNA and 24 8S rRNA), 38 micro-RNA hairpins and 146 other ncRNAs. Using the same annotation pipeline, we compared the abundance of each ncRNA family in parasitic (*Ascaris suum, Bursaphelenchus xylophilus, Brugia malayi, S. carpocasae, M. incognita, M. hapla and Heterorhabditis bacteriophora*) and free-living (*Panagrellus redivivus, Pristionchus pacificus, C. remanei and C. elegans*) nematodes (Supplementary Table S5). By comparing the average number of elements in each family between parasitic and free-living nematodes, we derived a simple metric to decide if a family had a tendency to be enriched in one of the two lifestyles (see Materials and Methods). The families enriched in parasitic nematodes are, ACEA_U3 (a snoRNA), SeC (a tRNA), mir-100/mir-10, mir-227, mir-2b, mir-2444, and mir-4455 (microRNAs), all of which have at least twice the number of elements on average in the parasites (Supplementary Fig. S2). Although the correlation between these families and parasitism needs to be further investigated, this is a first indication that these ncRNA families might have a functional role in the pathogenic lifestyle.

To complement these bioinformatic predictions we performed small RNA-seq of *S. carpocapsae* with and without induction with insect hemolymph. We obtained a total of 42.8 million reads from 6 libraries. Less than 10% of the cleaned reads failed to map to the genome (see Materials and Methods), another indication that the genome assembly is very complete. We used two of the most popular tools to annotate known and novel microRNAs using small RNA sequencing data: miRDeep[17] and ShortStack[18]. Both tools coincided in predicting 100 miRNAs, while

GO-ID	Term	Sc	As	Bm	Bx	Di	Hb	Ll	Mh	Mi	Ov	Sr	Ts	#Sp
GO:0005524	ATP binding	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	12
GO:0003743	translation initiation factor activity	↑	↑	↑		↑	↑	↑	↑	↑	↑	↑	↑	11
GO:0003964	RNA-directed DNA polymerase activity	↑	↑	↑	↑	↑	↑		↑	↑	↑	↑	↑	11
GO:0006278	RNA-dependent DNA replication	↑	↑	↑	↑	↑	↑		↑	↑	↑	↑	↑	11
GO:0015074	DNA integration	↑		↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	11
GO:0034754	cellular hormone metabolic process	↑	↑			↑		↑	↑	↑	↑	↑	↑	9
GO:0019915	lipid storage	↑	↑		↑	↑		↑	↑				↑	7
GO:0008284	positive regulation of cell proliferation	↑	↑				↑	↑			↑	↑		6
GO:0004190	aspartic-type endopeptidase activity	↑			↑				↑	↑		↑	↑	6
GO:0006446	regulation of translational initiation	↑	↑	↑				↑						4
GO:0006744	ubiquinone biosynthetic process	↑		↑		↑					↑			4
GO:0008270	zinc ion binding	↑		↑						↑			↑	4
GO:0045333	cellular respiration	↑		↑		↑					↑			4
GO:0055114	oxidation-reduction process	↑					↑		↑	↑				4
GO:0006913	nucleocytoplasmic transport	↑			↑							↑		3
GO:0015992	proton transport	↑			↑				↑					3
GO:0046339	diacylglycerol metabolic process	↑			↑					↑				3
GO:0004180	carboxypeptidase activity	↑				↑						↑		3
GO:0006886	intracellular protein transport	↑					↑					↑		3
GO:0009792	embryo development ending in birth or egg hatching	↑					↑					↑		3
GO:0006821	chloride transport	↑							↑			↑		3
GO:0004252	serine-type endopeptidase activity	↑								↑			↑	3

Table 3. Enriched GO terms in the genome of *Steinernema carpocapsae* and in at least two other pathogenic species but not in the free-living nematode *Pristionchus pacificus*, as compared to the free-living nematode *Caenorhabditis elegans*. *Sc, Steinernema carpocapsae*; As, *Ascaris summ*; Bm, *Brugia malayi*; Bx, *Bursaphelenchus xylophilus*; Di, *Dirofilaria immitis*; Hb, *Heterorhabditis bacteriophora*; Ll, *Loa loa*; Mh, *Meloidogyne hapla*; Mi, *M. incognita*; Ov, *Onchocerca volvulus*; Sr, *Strongyloides ratti*; Ts, *Taenia solium*; ↑, enriched term in that species (and in *S. carpocapsae* but not in *P. pacificus*) as compared with *C. elegans*.

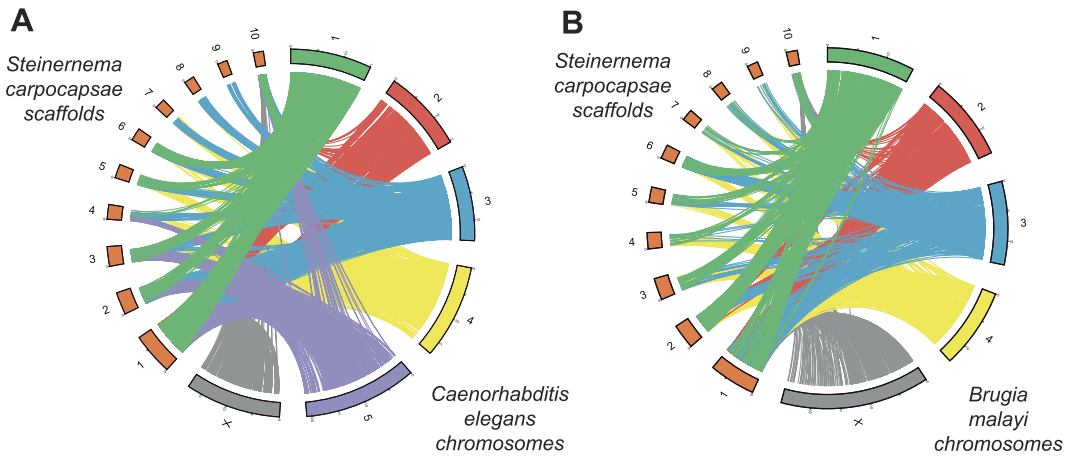


Figure 2. Schematic representation of shared sequences between *Steinernema carpocasae* and (**A**) *Caenorhabidits elegans*, or (**B**) *Brugia malayi*, both based in HSPs (E-value < 1e-6).

miRDeep predicted an additional 162, and ShortStack 25 more, giving a total of 287 miRNA hairpins, each with a potential 5′ and 3′ mature product (Supplementary Table S6). These predictions followed an expected length distribution, with a dominant peak centred at 22 nucleotides. Of the sRNA sequencing reads of 20–24 nucleotides that mapped to the genome, 83% overlapped with the 287 miRNA hairpins (Supplementary Fig. S3). Interestingly, we detected a large number of novel miRNA genes, since only 25 out of the 287 predicted miRNA hairpins correspond to known miRNAs according to homology searches. Although the majority of the conserved miRNAs tend to have high expression in our experiments, half of the 20 most highly expressed miRNA predictions correspond to novel sequences (Supplementary Table S7). This confirms the great diversity of ncRNA genes that are species or

Steinernema carpocapsae											
	Scaffold	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**
	Total genes	**1761**	**864**	**638**	**333**	**332**	**540**	**406**	**409**	**273**	**399**
Ce (%)	Chr1	5.91	5.90	4.23	1.20	4.82	55.37	3.69	6.60	5.13	6.77
	Chr2	30.44	7.06	6.90	4.20	23.49	7.96	6.65	5.87	6.59	9.27
	Chr3	5.79	37.73	3.76	5.11	5.42	5.19	5.91	7.09	6.96	9.02
	Chr4	10.96	8.22	7.99	4.20	24.40	7.04	7.14	8.80	44.32	8.02
	Chr5	6.93	6.71	39.34	23.12	5.12	6.30	5.42	5.87	5.13	9.77
	ChrX	26.69	4.05	5.33	4.20	4.22	5.19	18.23	17.85	5.49	27.82
	Without match	13.29	30.32	32.45	57.96	32.53	12.96	52.96	47.92	26.37	29.32
Bm (%)	Chr1	4.09	37.96	4.08	3.30	3.92	5.19	2.96	1.71	5.86	3.26
	Chr2	27.77	5.90	4.86	4.20	21.39	4.81	5.42	6.60	5.49	4.76
	Chr3	3.92	2.78	2.35	1.80	3.61	55.93	2.22	2.69	2.93	2.76
	Chr4	2.90	5.90	37.30	18.92	3.01	5.00	3.45	3.91	4.03	3.26
	ChrX	26.35	7.41	9.40	6.91	19.28	8.52	22.17	21.27	36.63	35.09
	Without match	34.98	40.05	42.01	64.86	48.80	20.56	63.79	63.81	45.05	50.88

Table 4. Percentage of genes located in single chromosomes (Chr) of *Caenorhabidtis elegans* (Ce, above) or *Brugia malayi* (Bm, below) that match genes located in single scaffolds of *Steinernema carpocapsae*.

lineage specific, and highlights the importance of using experimental data when annotating genomes, particularly for species that are distant to well annotated model organisms.

We were also interested to see if any of the microRNAs that we detected changed their expression in response to insect haemolymph. None of the miRNAs showed a significant decrease in expression, but five increased their expression after hemolymph induction (Supplementary Table S7 and Supplementary Fig. S4). These miRNAs were miR-84–3p, miR-84-5p, miR-31-3p, let-7-5p and Cluster_21397_3p (a new prediction with no similarity to known miRNAs). Interestingly, the induced miRNAs included miR-84 and let-7, members of the let-7 family of miRNAs that are important players during development. In *Caenorhabditis elegans*, double mutants of miR-84 and miR-48 (another let-7 family member) show a delayed moulting phenotype and accumulate a double cuticle[19]. This is interesting because *S. carpocapsae* infective juveniles have a double cuticle that is lost upon entering the insect host[8]. The up-regulation of miR-84 and let-7 could thus be involved in the moulting process, triggered by the contact with insect hemolymph.

Differentially expressed proteins. During infection, the nematodes first invade the insect intestine and then cross the intestinal wall by expressing putative effectors that facilitate parasite penetration to the hemocoel, where they continue to counteract insect defences[10,20]. Therefore, we were interested in comparing the soluble proteins from Infective Juveniles (IJs) induced with either insect intestines or insect hemolymph, against non-induced controls. We opted for a detection strategy combining shotgun proteomics strategy, with two-dimensional electrophoresis (2DE), and SDS-PAGE, that resulted in the identification of 7,527 proteins. By eliminating duplicates (proteins that were detected more than once), we obtained 3,773 non-redundant proteins (Supplementary Table S2). Among the non-redundant proteins, 1,625 were expressed in the three conditions, 155 were only expressed under both hemolymph and intestine induction, and 349 were expressed in the control and one other condition. In addition, 489 proteins were exclusively expressed in nematodes induced with intestine, 510 in those induced with hemolymph, and 645 in the non-induced controls (Fig. 3). This suggests that specific activities occur at different stages during the pathogenic process. Proteins expressed specifically in the induced conditions, were associated with functional categories (GO terms) using Blast2GO (Supplementary Figs S5 and S6). We found four GO terms enriched among the 489 proteins expressed exclusively in the intestine-induced sample, 10 GO terms in the 510 proteins of the hemolymph-induced sample, and 8 GO terms in the 1,154 combined proteins from both induction conditions (Fisher's exact test, $p < 0.05$); in all cases in relation to the untreated control (Supplementary Table S8). One of the differentially expressed proteins was a transthyretin-like protein (TLP). Unlike transthyretins, which are known for transporting thyroxine and related molecules, the function of TLPs is not well understood[21]. They seem to have a role in the uricase reaction pathway as 5-hydroxyisourate hydrolases[22]. Their abundance in parasitic nematodes and more specifically their expression during parasitic stages, suggests an involvement of TLPs in parasitism[22,23]. We also found other differentially expressed peptides in the parasitic stages of *S. carpocapsae* that, being found in other parasites, could be involved in the infective processes (Supplementary Table S2).

Proteases and excretory/secretory proteins. We found all types of proteases in the genome of *S. carpocapsae*. Proportions of most of them (Aspartic, Cysteine, and Threonine) are similar to the proportions found in free-living (*C. elegans*), necromeric (*P. pacificus*), and parasitic (*B. malayi, Strongyloides ratti*) nematodes. However, the amount of serine proteases in *S. carpocapsae* and *P. pacificus* genomes is higher than in other nematodes (39% and 33.2% of all proteases, respectively); while in *S. carpocapsae* the percentage of metalloproteases, although high (31%), is the lowest in the comparison (the highest is in the *Strongyloides ratti* genome with 52%).

Excretory/secretory (ES) products are complex mixtures of hundreds of different proteins that are thought to have important roles in the life cycle of a parasite and during host-parasite interactions[24]. A total of 1,421

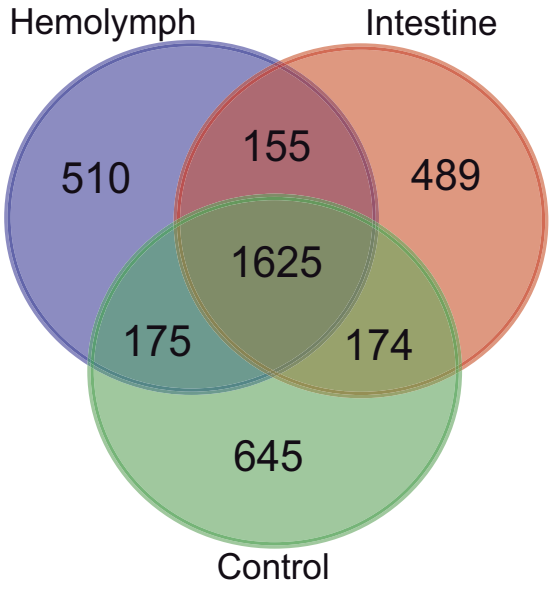


Figure 3. Non-redundant soluble proteins expressed after induction of *Steinernema carpocapsae* IJs with insect intestines, hemolymph or non-induced control.

putative excreted proteins were predicted in the genome (see Methods), including various families of proteases, protease inhibitors, cuticular collagens, and C-type lectins, as well as putative signalling molecules such as warthog, ground and ground-like proteins (Supplementary Table S9). Some ES proteins are predicted to be involved in immuno-evasion; such as collagen[25], whereas others play crucial roles in the suppression of host immune responses by mimicking host molecules, such as C-lectin[26,27]. Furthermore, C-lectins have been found to be upregulated in *Ancylostoma ceylanicum* at the onset of heavy blood feeding from the host[28]. We also found three putative copies of parasitic stage specific protein 1, a protein without known domains, which is present in several parasitic nematodes and is expressed during the transition to the parasitic lifestyle in *Haemonchus contortus*[29]. Other interesting putative secreted peptides in the *S. carpocapsae* genome were lipases, saposins, and transthyretin-like protein, all of which are expressed during early parasitic stages in *H. contortus*[29]. Serine proteases, including Sc-SP-1 and Sc-SP-3, can mediate the invasion or apoptosis of host cells[10,20]. Astacin metalloprotease is one of the effector molecules involved in tissue invasion of parasitic nematodes[30]. The family of papain-type aspartic and cysteine proteases are thought to have the same role in invertebrate digestion as trypsin in vertebrates[31]. Therefore, it is possible that the dependence of *S. carpocapsae* on aspartic protease activities is related to the digestion of nutrients. Cysteine proteases are involved in digestive processes or moulting and cuticle renewal in free-living and parasitic nematodes[32,33].

Orthologous proteins. We compared 7,724 orthologous groups of proteins among several species with different lifestyles. We found 318 orthologous groups that are absent in non-pathogenic species (*Caenorhabditis angaria, C. remanei, C. briggsae, C. japonica, C. elegans, and Pristionchus pacificus*) but present in *S. carpocapsae* and in at least another pathogenic nematode (Supplementary Table S10). The annotations of these orthologues revealed enrichment of protein functions with possible associations with parasitism, such as serine proteases and other terms related to degradation and binding (Table 5). We also found 134 additional groups from the orthoMCL-DB (version 5) database that are present in the genome of *S. carpocapsae* but not in the other tested species (Supplementary Table S11).

Positive selection. Because of the co-evolutionary arms-race relationship between hosts and their pathogens, genes involved in their interaction are expected to evolve under positive selection[34], potentially resulting in specific genomic signatures associated with their lifestyles[35]. We used the branch-sites test of positive selection[36,37] to analyse 2,034 orthologous genes in three species of Clade IV nematodes (as defined in ref. 38). This test is based on a maximum likelihood estimation of the nucleotide nonsynonymous and synonymous substitutions rates. The ratio of nonsynonymous to synonymous rates (ω) can be used to identify purifying selection ($\omega < 1$), neutral evolution ($\omega = 1$), or positive selection ($\omega > 1$), assessing the significance with a Likelihood Ratio Test (LRT)[36]. We found 83 genes with sites evolving under positive selection ($\omega > 1$, LRT, $p < 0.05$; 14 of which had an FDR < 0.1) in *S. carpocapsae* (Table 6). Among the 83 genes, 23 GO terms were significantly enriched (Supplementary Table S12) when compared to the genes with no sites under positive selection (1,951 genes) (Fisher's exact test, $p < 0.01$). Although we found more genes (95) with sites evolving under positive selection in the free-living nematode *Panagrellus redivivus*, there were no enriched GO terms among them (even with an alpha value of 0.05), indicating that the consequence of an arms-race relationship is better characterized by positive selection preferentially altering genes of specific functions than just increasing the number of positively

GO-ID	Term	Category	FDR	P-Value
GO:0045449	regulation of transcription, DNA-dependent	P	4.30E-16	8.08E-20
GO:0004252	serine-type endopeptidase activity	F	7.73E-15	2.90E-18
GO:0005667	transcription factor complex	C	8.03E-12	4.52E-15
GO:0008236	serine-type peptidase activity	F	1.25E-11	1.17E-14
GO:0017171	serine hydrolase activity	F	1.25E-11	1.17E-14
GO:0045941	positive regulation of transcription, DNA-dependent	P	2.16E-05	2.44E-08
GO:0004175	endopeptidase activity	F	1.78E-04	2.34E-07
GO:0043234	protein complex	C	3.18E-03	5.09E-06
GO:0005515	protein binding	F	3.18E-03	5.37E-06
GO:0045935	positive regulation of nucleobase-containing compound metabolic process	P	4.95E-03	1.11E-05
GO:0051254	positive regulation of RNA metabolic process	P	4.95E-03	1.11E-05
GO:0051173	positive regulation of nitrogen compound metabolic process	P	4.95E-03	1.11E-05
GO:0008233	peptidase activity	F	8.30E-03	2.02E-05
GO:0003713	transcription coactivator activity	F	8.41E-03	2.21E-05
GO:0010628	positive regulation of gene expression	P	1.00E-02	2.98E-05
GO:0043170	macromolecule metabolic process	P	1.00E-02	3.02E-05
GO:0070011	peptidase activity, acting on L-amino acid peptides	F	1.01E-02	3.40E-05
GO:0006508	proteolysis	P	1.01E-02	3.42E-05
GO:0010557	positive regulation of macromolecule biosynthetic process	P	3.65E-02	1.49E-04
GO:0045298	tubulin complex	C	3.65E-02	1.58E-04
GO:0033202	DNA helicase complex	C	3.65E-02	1.58E-04
GO:0031011	Ino80 complex	C	3.65E-02	1.58E-04
GO:0097346	INO80-type complex	C	3.65E-02	1.58E-04

Table 5. Enrichment of Gene Ontology (GO) terms of orthologs abscent in non-pathogenic nematodes (*Caenorhabditis angaria, C. remanei, C. briggsae, C. japonica, C. elegans, and Pristionchus pacificus*) and present in *S. carpocapsae* and at least another parasitic nematode compared to *C. elegans* GO terms.

Orthologue genes analysed by the Branch-site test	N = 2,034		
Tested branch	Sc	Pr	Sr
Genes with sites under positive selection ($\omega > 1$, LRT, $p < 0.05$)	83 (4.08%)	95 (4.67%)	8 (0.39%)
Average proportion of sites under positive selection per gene (s.d.)	6.49% (0.067)	7.47% (0.093)	17.31% (0.187)

Table 6. Genes with sites evolving under positive selection in *Steinernema carpocapsae* (Sc), *Panagrellus redivivus* (Pr), and *Strongyloides ratti* (Sr).

selected genes. Although we would need to increase the number of analysed genes to increase the power of these analyses, the initial results suggest a new perspective to the co-evolutionary arms-race theories.

Phylogenetic analysis. To explore the phylogenetic relationships of *S. carpocapsae*, we reconstructed a phylogeny using 245 proteins from strictly 1-1 orthologous genes from nine nematode species. According to Blaxter *et al.*[38], *Steinernema* is phylogenetically closer to *Strongyloides* than to *Caenorhabditis,* as inferred from a tree reconstructed using the small subunit ribosomal DNA (18S) sequences from 53 nematode species. A similar result was obtained in a more extensive analysis using 339 18S sequences[39]. However, Montiel *et al.*[40] found *Steinernema* to be closer to *Caenorhabditis* than to *Strongyloides* using complete mtDNA sequences. Although this discrepancy may result from differential reproductive strategies and/or differential selective pressures acting on nuclear and mitochondrial genes[40], an analysis of large subunit ribosomal DNA sequences (28S) also showed *Steinernema* to be closer to *Caenorhabditis*[41]. Defining these relationships is relevant because if *Steinernema* is phylogenetically closer to *Strongyloides*, it could be used as a more general model for parasitism, with implications for human health. *Steinernema* is more tractable than *Strongyloides* because it does not require a vertebrate host to reproduce in the laboratory. Our new phylogenetic analysis supports *Steinernema* being closer to *Strongyloides* than to *Caenorhabditis* (Fig. 4). In addition, its basal position in relation to *Strongyloides*, gives support to the hypothesis that this vertebrate parasite originated by host shifting from an entomopathogenic ancestor[12], in this case *Steinernema*.

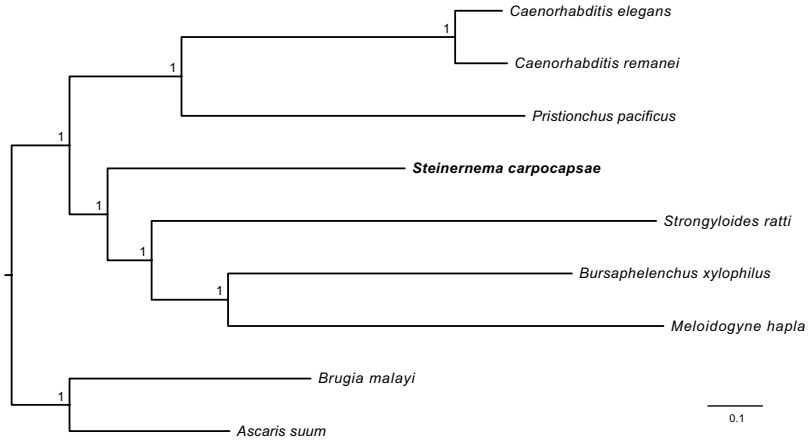


Figure 4. Bayesian phylogenetic tree reconstructed from the concatenated alignment of 245 orthologous proteins of nine nematode species. Numbers in branches are posterior probabilities.

Gene functions enriched at several levels. To assess how the pathogenic lifestyle is affecting specific gene functions at different levels, we compared the GO terms enriched in the genome of *S. carpocapsae* when compared to *C. elegans* (Supplementary Table S3), with those in differentially expressed proteins due to hemolymph or intestine induction (Supplementary Table S8), and with those in genes with putative sites evolving under positive selection (Supplementary Table S12). One GO term was enriched in the genome and in differentially expressed proteins (transcription factor activity – sequence-specific DNA binding); and two GO terms in both the genome and genes under positive selection (macromolecular complex and ribonucleoprotein complex). No GO terms were shared between the three analyses. However, one protein, ATAD-3 (ATPase family AAA domain-containing protein 3), is associated with 33 enriched GO terms in the annotated sequences, and two additional enriched GO terms from genes with evidence of positive selection (Supplementary Table S13). This protein is differentially expressed in nematodes induced with insect tissues (intestines) and presents amino acid sites evolving under positive selection. Functions or proteins shared between these analyses might reveal relevant effects of the pathogenic lifestyle in the genome. A deeper analysis of positive selection (i.e. including more orthologous genes or conducting population genetic analyses) could expand the number of shared genes, which should be good candidates for further studies. In this case we have identified a putative homolog of *C. elegans*' ATAD-3. Its deficiency in *C. elegans* causes early larval arrest, gonadal dysfunction, and embryonic lethality. It is also associated with defects in organellar structure and mtDNA depletion[42,43], suggesting that ATAD-3 is important for increased mitochondrial activity during the transition to later larval stages[42]. *S. carpocapsae* needs to go through developmental changes to establish itself in the insect body during the pathogenic process, which might explain the relevance of this, and probably other mitochondria-related genes, in nematode parasitism. For example, the defective mitochondrial respiration family member protein 1, with functions in regulation of growth rate, was differentially expressed in nematodes induced with insect tissues (hemolymph) and presented evidence of positive selection. Mitochondria have been identified as important contributors to the virulence of fungal pathogens[44], and it has been previously hypothesised that differential selective constraints in mitochondrial genes might explain discrepancies between nuclear and mitochondrial gene phylogenies in nematodes[40]. In addition, depletion of ATAD-3 in *C. elegans* resulted in reduced intestinal fat storage[42], and it would be interesting to explore if fat metabolism might also be relevant in nematode parasitism.

Conclusion

Our genomic analyses of *S. carpocapsae* confirm a role in pathogenicity beyond simply vectoring the symbiotic bacteria. *S. carpocapsae* shares orthologous genes with other parasitic nematodes that are absent in the free-living nematode *C. elegans*, it encodes ncRNA families that are enriched in parasites, and presents putative proteins associated with functions related to parasitism and pathogenesis. Until now, the best examples of positive selection in genes related to host-pathogen interactions were pathogen effectors and genes of the host immune and defence systems[34]. Our analyses indicate that positive selection can also alter genes belonging to other functional categories, such as metabolism and development, adding a new aspect to the arms-race co-evolutionary theories. Through a comprehensive analysis, we identified a protein, ATAD-3, suggesting a relevant role for mitochondria during the evolution of nematode parasitism that warrants further investigation. We provide additional evidence for the phylogenetically relatedness of *S. carpocapsae* to *Strongyloides*, making this high-quality genome valuable for comparative studies with potential implications for human health. Our genome also represents a useful resource to aid ongoing efforts towards the genetic improvement of entomopathogens as biological control agents as well as to better understand host-parasite interactions in nematodes.

Materials and Methods

Organisms, maintenance and storage. *Steinernema carpocapsae* strain Breton was obtained from Nelson Simões, and cultured using *in vitro* methods. Nematodes were grown using a modified protocol for mass production in artificial medium according to ref. 45, as well as in small-scale, on plates containing Fortified Lipid

Agar (FLA) prepared with 1.6% TSB (nutrient broth), 1% vegetable oil, 1.2% bacteriological agar, and 5% yeast extract (modified from ref. 46). A near isogenic line of *S. carpocapsae* strain Breton was generated by reproduction of single couples of brother and sister for 12 generations (F12). This produces ~96% homozygosis[47].

DNA isolation, sequencing and quality control. Total genomic DNA was isolated from the nuclei[48] using the phenol/chloroform extraction protocol described by Sambrook *et al.*[49]. Total DNA yield and integrity was measured with a 2100 Bioanalyzer (Agilent) using an Expert High Sensitivity DNA chip. Three high-quality libraries for Next Generation Sequencing (NGS) were prepared following manufacturer's instructions. One shotgun 454 library was sequenced in three 454 FLX runs. One 454 paired-end library with 8-kb inserts was sequenced in two 454 FLX runs. Finally, a SOLiD shotgun library was tagged and sequenced, along with a different library (from other organisms), in a lane of SOLiD 5500xl, equivalent to half a lane of SOLiD sequencing. Low-quality sequences, base-calling duplicates and adapters were removed from all the sequence data (see below).

Genomic assembly and filtering. All DNA-sequence reads were filtered to remove contamination of the endosymbiotic bacteria *Xenorhabdus nematophila* (Xn). A genomic dataset was created by adding the published genome of Xn strain ATCC 19061 to the unpublished genome of the Xn strain isolated from the *S. carpocapsae* strain Breton nematodes, produced in our laboratory. The dataset was used for contamination screening and filtering using GS Assembler 2.7.

Raw standard flowgram format (sff) files coming from 454 platforms were assembled using GS Assembler 2.7 with the default trimming parameters. Basespace reads coming from the SOLiD platform using the Exact Call Chemistry Module (that allows conversion from colour to basespace) were filtered to remove PCR clonal repeats as well as reads with ambiguous bases. Subsequently, sequences were filtered based on Phred quality values. Bases below Phred18 were removed from 3′ ends and only reads longer than 20 bp were kept. Filtered data were assembled into contigs using GS Assembler 2.7, and joined into scaffolds using the paired-end data.

GC-content was estimated from the scaffolds using 10-kb non-overlapping sliding windows, and GC-bias was assessed based on a frequency distribution of these data. To evaluate the completeness of the genome assembly, we followed two strategies. RNA-seq sequences representing all different stages and diverse culture conditions of *S. carpocapsae* were mapped to the final assembly using Newbler (GS Reference Mapper v.2.7). In addition, we analysed the completeness of 248 ultra-conserved core eukaryotic genes[13]. We expect a complete genome will contain a higher number of complete ultra-conserved genes.

Estimation of genome size. Genome size was estimated from the genomic assembly using GS Assembler 2.7 and corroborated through flow cytometry of the isolated cellular nuclei of *S. carpocapsae* using the nuclei of *C. elegans* strain N2 (genome size approx. 100 Mb[50]) as a size control. Nuclei were stained with CyStain® UV Ploidy (Partec 05-5001), and fluorescence was detected at $\lambda \leq 420$ nm and quantified using a PARTECPAII (Partec, Germany) flow cytometer with a mercury lamp (100 W UV light).

Assessment of repeat content. Following genome assembly, repeats were identified using a combination of homology-based comparisons (using RepeatMasker[51]) and a *de novo* approach (using RepeatModeler[52]).

Annotation of non-coding RNA. Covariance models from Rfam[53] were used to scan the genomes using Infernal software[54]. In addition, tRNAs were predicted using tRNAscan-SE[55] and rRNAs were predicted using RNAmmer[56]. Finally, microRNA precursor sequences (miRNA hairpins) were located using MapMi[57], using all mature miRNA sequences from miRBase 21[58] as input. MapMi results were filtered selecting only microRNA precursor sequences with score ≥ 30. Results from Rfam, tRNAscan-SE, RNAmmer and MapMi were processed within R (R: A Language and Environment for Statistical Computing; http://www.r-project.org), using 'GenomicFeatures' and 'rtracklayer' packages[59,60].

To compare ncRNA families present in parasitic (*Ascaris suum, Bursaphelenchus xylophilus, Brugia malayi, S. carpocasae, Meloidogyne incognita, M. hapla, and Heterorhabditis bacteriophora*) and free-living (*Panagrellus redivivus, Pristionchus pacificus, Caenorhabditis remanei, and C. elegans*) nematodes, the average number of genes belonging to each ncRNA family, were calculated for each group. Families with at least twice the average number of genes in the parasitic compared to free-living group were selected.

RNA isolation, sequencing and assembly. Total RNA was extracted from a pool of individuals from all lifecycle stages (eggs at different stages, L1, L2, L3, IJ, L4, and adults), cultured *in vivo* infecting *Galleria mellonella* and *Tenebrio molitor* larvae as described[61], as well as all lifecycle stages cultured *in vitro* using the methods described in the Organisms, maintenance and storage section[45,46]. The final pool consisted of approximately 3 mg of individuals from each stage/condition. RNA was extracted using TRIzol (Invitrogen) according to the manufacturer's instructions with an additional step using Qiagen RNAeasy Mini Elute Clean up columns and buffers to clean and concentrate the RNA.

An Illumina paired-end library and a 454 library were generated for RNA-seq, which were run on a full plate of MiSeq and a partial plate of a 454 FLX+, respectively. RNA-seq reads were quality filtered and mapped to the repeat-masked genome using Newbler gsmapper. Read alignments were provided to Trinity[62] (r2013-02-25) as a coordinate-sorted bam file. Trinity was used to assemble the aligned reads. The Trinity-reconstructed transcripts were aligned and assembled using the PASA2[63] (r20130425 beta) pipeline.

For small RNA (sRNA) sequencing, nematodes were induced for 2 hour with hemolymph of *Galleria mellonella* and with buffer as control[64]. Nematodes were grinded under liquid nitrogen and RNA was extracted with

Trizol according to the manufacturer's instructions. Six sRNA-Seq tagged libraries were prepared from three replicates of each condition, which were run in an Illumina HiSeq lane.

Processing small RNA sequencing results. All sRNA-Seq libraries were 3′-adaptor trimmed using the reaper tool from Kraken[65]. After trimming, reads between 18 and 36 nucleotides were mapped to the genome using ShortStack 3.3[18], setting the maximum number of mismatches to 1, no stich, multi-mappers guided by unique-mappers and removing reads that mapped to more than 101 locations. The raw and processed sRNA-Seq results were deposited in GEO (http://www.ncbi.nlm.nih.gov/geo), under accession GSE85256.

Predicting expressed microRNA loci with miRDeep and ShortStack. When using ShortStack to annotate, we set the Dicer minimum and maximum size parameters to 18 and 36, minimum alignment coverage (mincov) to 5 and maximum distance to merge clusters (pad) to 50. To improve the predicted microRNA producing loci, we used MirDeep2[17]. The mapper.pl and miRDeep2.pl modules were used to identify known and novel microRNAs. Reads between 18 and 36 nucleotides were used, with a maximum number of mismatches of 1, and 101 maximum number of locations for multi-mapping reads. The minimum alignment coverage (-a) was set to 5, the maximum number of precursors to analyze was set to 1000, and all the mature sequences from miRBase 21 were provided. Although both programs take small RNA sequencing reads mapped to a genome to predict microRNA loci, they produce slightly different results. They coincided in predicting 100 miRNAs, while miRDeep predicted an additional 162 and ShortStack 25 more.

Differential expression analysis of microRNAs. To focus the differential expression analysis on miRNAs, only reads in the 20–24 nucleotide length range were considered. After mapping these to the genome, on average 83% fell within miRNA hairpin and 63% within mature miRNA coordinates. For miRNA quantification, the featureCounts function of the Rsubread R package[66] was used, asking for a minimum overlap of one nucleotide to any of the mature miRNA annotations. For differential expression analysis, the edgeR package was used[67]. miRNAs with less than 3 counts-per-million in at least 3 libraries were removed, leaving 302 out of the 574 annotated mature miRNAs. The trimmed mean of M-values was chosen as normalization method[68]. Genewise data dispersion was estimated with the function estimateGLMTagwiseDisp, which uses an empirical Bayes strategy[67]. Differentially expressed miRNAs were determined with a generalized linear model and gene-wise likelihood ratio tests. A False Discovery Rate threshold of 0.1 was selected to consider a miRNA to be significantly differentially expressed. According to this threshold, five miRNAs were overexpressed in response to hemolymph treatment and none were down regulated. Cluster_19164 was identified as miR-84 by a manual sequence search on the miRBase website[58].

Gene prediction and synteny. The *S. carpocapsae* protein-coding gene set was inferred using *de novo*, homology- and evidence-based approaches (Supplementary Fig. S7). *De novo* gene prediction was performed on a repeat-masked genome using Augustus[14]. Training models were generated using hints from a compilation of *S. carpocapsae* gene structures (CEGMA[13] [v2.4.010312] predictions, PASA assemblies from our RNA-seq data, and 2,269 publicly available ESTs from GeneBank). The homology-based prediction was conducted with Augustus algorithms for *C. elegans* and *Brugia malayi*. Synteny was assessed on scaffolds >1 Mb using pairwise alignments with E-value $< 10^{-6}$ and homologous regions were visualized using CIRCOS[69]. Macro-synteny was analysed using the SynFind and SynMap tools from CoGe[70,71].

Functional annotation of coding genes. Following the prediction of the protein-coding gene set, we conducted high-stringency BLASTp homology searches (E-value $\leq 10^{-5}$) against the NCBI non-redundant protein database. Functional annotation was performed using Blast2GO[72]. Gene ontology categories were summarized and standardized to level 2 and level 3 terms, defined using the GOslim hierarchy[73]. For the secretome prediction, the signal peptide was predicted by SignalP 4.0[74] and Phobious[75] employing both Hidden Markov Models and Neural Networks. Proteins were then filtered for the presence of transmembrane regions using THMMN[76] and Phobious[75]. Subcellular localizations were identified using TargetP ($\geq$95% specificity)[77] and WolfPSORT[78] (score $\geq$30). Proteases and protease inhibitors were identified by homology searches to the MEROPS database[79].

Orthologous proteins. Genes from different nematode species (*Caenorhabditis angaria, C. briggsae, C. elegans, C. japonica, C. remanei, Pristionchus pacificus, Ascaris suum, Brugia malayi, Bursaphelenchus xylophilus, Heterorhabditis bacteriophora, Haemonchus contortus, Loa loa, Meloidogyne hapla, M. incognita, Onchocerca volvulus, Panagrellus redivivus, S. carpocasae, Strongyloides ratti, Trichinella spiralis, and Wuchereria bancrofti*) were assigned to OrthoMCL[80] orthologous groups, and the presence or absence of groups was compared among the different species using ad hoc scripts from scriptome (http://archive.sysbio.harvard.edu/csb/resources/computational/scriptome/UNIX/).

Additional bioinformatics analyses and use of software. Data analysis was conducted in a UNIX environment or Microsoft Excel 2007 using standard commands. Bioinformatic scripts required to facilitate data analysis were designed using bash, GNU coreutils, Perl, and Python.

Proteomic analysis. *Sample preparation.* Nematodes were induced as described[64] with slight modifications. A pool of approximately 25,000 nematodes were induced with either hemolymph or intestine of *Galleria mellonella*. To obtain hemolymph we grinded *G. mellonella* larvae in liquid nitrogen, added a volume of cold

Tyrode's solution (NaCl 0.8%, KCl 0.02%, $CaCl_2$ 0.02%, $MgCl_2$, 0.02%, NaH_2PO_4, 0.005%, $NaHCO_3$, 0.1%, and glucose, 0.1%) and sonicated at a frequency of 20 kHz. Then we centrifuged at 1700 rcf for 15 min at 4 °C, obtaining three phases, of which the middle one corresponded to hemolymph. Intestines were dissected from insect larvae, collected on a watch glass and rinsed several times with sterile saline solution (NaCl 0.8%) to eliminate any trace of hemolymph or other remains. Nematode infecting juveniles (IJs) were superficially disinfected with 2% sodium hypochlorite during 10 min, and rinsed three times with sterilized water. According to our experience, this treatment is enough to kill all surface bacteria and fungi from the nematodes while keeping them viable. Washed nematodes were then transferred to a 90 × 15 mm Petri dish containing 7 ml of Tyrode's solution with 10% of *G. mellonella* hemolymph (v/v) or 10% of intestines (w/v), and 1% Nalidixic acid, to avoid contamination. Previous experience in the Simões lab has shown that 10% of hemolymph is needed to induce recovery of the IJs and to allow them to complete their life cycle. The same concentration of intestines was used as a first approximation to understand the effects of insect intestines on protein expression. Different pools of nematodes were incubated under agitation (40 rpm) at 25 °C for 1, 2, 4 and 8 hours, and analysed separately. To determine these time points we followed the infection process *in vivo*, by conducting dissections at regular intervals, to understand the kinetics of the infection. We observed that after entering the intestine, it took the nematodes one hour to start traversing the intestine wall and at two hours, most of them were in the hemocoel. We included two additional time points to capture proteins expressed lately in the infection process; stopping at 8 hours because at this point the nematodes start to release the symbiotic bacteria[81]. Nematodes without induction were used as a negative control. Nematodes were grinded in liquid nitrogen and suspended in lysis buffer (7 M urea, 2 M thiourea, 3% CHAPS) with protease inhibitor mix GE and sonicated at a frequency of 20 kHz. After centrifugation (1 min at 16,000 rcf) and filtering (0.45 μm Millex-HV PVDF, Millipore), the supernatant was precipitated with the 2D Clean-up kit (GE Healthcare), resuspended in DeStreak solution (GE Healthcare Cat. No. 17600318), and quantified with the Bradford Method[82] (BioRad Protein Assay Dye Cat. No. 500-0006).

SDS-PAGE and 2D electrophoresis. Protein samples (80 μg) were loaded onto the SDS-PAGE and ran at 100 V in a vertical Mini-PROTEAN Tetra cell BioRad. Gels were stained with Coomassie blue. Isoelectric focusing was performed using GE Healthcare Immobiline strips pH 3-10 and DryStrip pH 4–7, in both cases of 7 cm of length. Isoelectric focusing was performed on a Multiphor II (GE Healthcare) using the conditions recommended by the manufacturer. The second dimension was run on 13% polyacrylamide gels.

Shotgun proteomics. Two hundred μl of the total protein extract was fractionated by isoelectric focusing in an IEF ZOOM® Fractionator system (Life Technologies) using the protocol described by the manufacturer. The protein pellet was passed through a reduction/alkylation process with urea (6 M), dithiothreitol (DTT) at a final concentration of 5 mM, and iodoacetamide (IAA) at a final concentration of 15 mM. Peptides were digested with trypsin (Promega) overnight at 37 °C and desalted with a Macro Spin Column (Nest Group). Thirty μg of protein from each fraction were then analyzed by LC-MS/MS in the Proteomics Facility of the UC Davis Genome Center. A Thermo Scientific Q Exactive Orbitrap MS spectrometer was used in conjunction with a Proxeon Easy-nLC II HPLC (Thermo Scientific) and a source Proxeon nanospray using a column 100 micron × 25 mm Magic C18 5U 100 Å reverse phase. The MS/MS spectra were acquired using the TOP15 method following the equipment manufacturer's instructions. All analyses were run in duplicates, including treatment samples and controls. ProteinPilot (v4.5), Mascot (v2.4), MaxQuant (v1.3.0.5), and Sequest (v1.3) software were used to identify peptides and proteins in each sample (see software references in Supplementary Table S2). In all cases, a tolerance in the mass measurement of 50 ppm in MS mode and 0.5 Da for MS/MS ions was used, with a significant threshold set to $p < 0.05$ and a confidence value ≥95%, with the exception of MaxQuant, in which the peptide mass tolerance was of 20 ppm, the fragment mass tolerance of 0.5 Da, and the confidence value was ≥99%. Modifications allowed were carbamidomethylation C (fixed), deamination NQ (variable), and oxidation M (variable).

Proteins detected in at least one sample replica and undetected in the two control replicates were considered as differentially expressed proteins. Proteins detected in at least one of the control replicates, but undetected in the two sample replicates were considered differentially suppressed proteins. Annotation and functional enrichment of differentially expressed proteins were performed with Blast2GO[72]. The Fisher Exact Test was used to compare the GO terms identified under the different induction conditions (hemolymph or gut) with the nematodes without induction.

Phylogenetic analysis. Protein sequences of all organisms were downloaded from WormBase (ftp.wormbase.org release WS241 29-Nov-2013). Orthologous genes for nine species were identified with OrthoMCL (v5)[80]. A total of 245 orthologous proteins were aligned with MUSCLE (v3.8.31)[83]. Phylogenetically informative blocks were recovered with Gblocks[84] and the best-fit evolutionary model for each aligned protein was predicted by ProtTest[85]. MrBayes[86] was used for phylogenetic reconstruction using concatenated alignments. Partitions were created by grouping proteins according to their best-fit model, i.e. each partition contained all the proteins evolving under the same model. A mixed model was applied to each partition, with different G, I, and F parameters and unlinking the model between partitions. To check for convergence, two runs with four chains each were performed. The analysis was run for 1,000,000 generations, and a burn-in of 25% was used. *Brugia malayi* and *Ascaris suum* were used to root the tree because these species were the most phylogenetically basal of the nine nematode species in the phylogenies obtained by both Blaxter *et al.*[38] and Nadler *et al.*[41].

Analysis of positive selection. We used protein-coding genes from *Panagrellus redivivus, Strongyloides ratti* and *Steinernema carpocapsae*, all nematodes from phylogenetic clade IV, according to Blaxter *et al.*[38]. All nucleic and amino acid sequences, except for *S. carpocapsae*, were downloaded from WormBase (ftp.wormbase.

org release WS241 29-Nov-2013). Orthologues proteins obtained with OrthoMCL (v5)[80] were aligned with ClustalW2 (v2.1)[87]. After selecting phylogenetically informative sites with Gblocks[84], and estimating the best-fit model with ProtTest[85], we reconstructed a consensus phylogenetic tree with PhyML (v.3.0)[88]. A nucleotide alignment based on the complete amino acid alignment was obtained with RevTrans (v1.4)[89] to preserve codon homology. The tree and the nucleotide alignments of each orthologous gene were used to assess signatures of natural selection with Codeml from the PAML package (v.4.6)[90], using the Branch-site model to identify genes with sites under positive selection. Annotation and functional enrichment in genes with positively selected sites were performed with Blast2GO[72].

References

1. Oerke, E.-C. Crop losses to pests. *The Journal of Agricultural Science* **144,** 31–43, doi: 10.1017/S0021859605005708 (2006).
2. Aktar, W., Sengupta, D. & Chowdhury, A. Impact of pesticides use in agriculture: their benefits and hazards. *Interdisciplinary Toxicology* **2,** 1–12 (2009).
3. Mukherjee, P. K., Horwitz, B. A., Herrera-Estrella, A., Schmoll, M. & Kenerley, C. M. Trichoderma research in the genome era. *Annual Review of Phytopathology* **51,** 105–129 (2013).
4. Lu, D., Baiocchi, T. & Dillman, A. R. Genomics of entomopathogenic nematodes and implications for pest control. *Trends in Parasitology* **32,** 588–598, doi: 10.1016/j.pt.2016.04.008 (2016).
5. Murfin, K. E. *et al.* Nematode-bacterium symbioses—cooperation and conflict revealed in the "Omics" age. *The Biological Bulletin* **223,** 85–102 (2012).
6. Thaler, J.-O., Duvic, B., Givaudan, A. & Boemare, N. Isolation and entomotoxic properties of the *Xenorhabdus nematophilus* F1 lecithinase. *Applied and Environmental Microbiology* **64,** 2367–2373 (1998).
7. Caldas, C., Cherqui, A., Pereira, A. & Simões, N. Purification and characterization of an extracellular protease from *Xenorhabdus nematophila* involved in insect immunosuppression. *Applied and Environmental Microbiology* **68,** 1297–1304, doi: 10.1128/aem.68.3.1297-1304.2002 (2002).
8. Herbert, E. E. & Goodrich-Blair, H. Friend and foe: the two faces of *Xenorhabdus nematophila*. *Nature Reviews Microbiology* **5,** 634–646 (2007).
9. Binda-Rossetti, S., Mastore, M., Protasoni, M. & Brivio, M. F. Effects of an entomopathogen nematode on the immune response of the insect pest red palm weevil: Focus on the host antimicrobial response. *J Invertebr Pathol* **133,** 110–119 (2016).
10. Toubarro, D. *et al.* An apoptosis-inducing serine protease secreted by the entomopathogenic nematode *Steinernema carpocapsae*. *International Journal for Parasitology* **39,** 1319–1330, doi: http://dx.doi.org/10.1016/j.ijpara.2009.04.013 (2009).
11. Dillman, A. R. *et al.* Comparative genomics of *Steinernema* reveals deeply conserved gene regulatory networks. *Genome Biology* **16,** 1–21 (2015).
12. Blaxter, M. & Koutsovoulos, G. The evolution of parasitism in Nematoda. *Parasitology* **142,** S26–S39 (2015).
13. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23,** 1061–1067, doi: 10.1093/bioinformatics/btm071 (2007).
14. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research* **33,** W465–W467 (2005).
15. Douzery, E. J., Snell, E. A., Bapteste, E., Delsuc, F. & Philippe, H. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *PNAS* **101,** 15386–15391 (2004).
16. Quentin, M., Abad, P. & Favery, B. Plant parasitic nematode effectors target host defense and nuclear functions to establish feeding cells. *Frontiers in Plant Science* **4,** 53, doi: 10.3389/fpls.2013.00053 (2013).
17. Friedländer, M. R. *et al.* Discovering microRNAs from deep sequencing data using miRDeep. *Nature biotechnology* **26,** 407–415 (2008).
18. Axtell, M. J. ShortStack: comprehensive annotation and quantification of small RNA genes. *Rna* **19,** 740–751 (2013).
19. Abbott, A. L. *et al.* The let-7 MicroRNA family members mir-48, mir-84, and mir-241 function together to regulate developmental timing in *Caenorhabditis elegans*. *Developmental cell* **9,** 403–414 (2005).
20. Toubarro, D. *et al.* Serine protease-mediated host invasion by the parasitic nematode *Steinernema carpocapsae*. *Journal of Biological Chemistry* **285,** 30666–30675 (2010).
21. Hennebry, S. C., Law, R. H., Richardson, S. J., Buckle, A. M. & Whisstock, J. C. The crystal structure of the transthyretin-like protein from *Salmonella dublin*, a prokaryote 5-hydroxyisourate hydrolase. *Journal of molecular biology* **359,** 1389–1399 (2006).
22. Lee, Y. *et al.* Transthyretin-related proteins function to facilitate the hydrolysis of 5-hydroxyisourate, the end product of the uricase reaction. *FEBS Lett* **579,** 4769–4774 (2005).
23. Furlanetto, C., Cardle, L., Brown, D. & Jones, J. Analysis of expressed sequence tags from the ectoparasitic nematode *Xiphinema index*. *Nematology* **7,** 95–104, doi: 10.1163/1568541054192180 (2005).
24. Britton, C. 18 Proteases of Nematodes: From Free-living to Parasite. *Parasitic Nematodes: Molecular Biology, Biochemistry and Immunology* 351 (2013).
25. Blaxter, M., Page, A., Rudin, W. & Maizels, R. Nematode surface coats: actively evading immunity. *Parasitology Today* **8,** 243–247 (1992).
26. Yoshida, A., Nagayasu, E., Horii, Y. & Maruyama, H. A novel C-type lectin identified by EST analysis in tissue migratory larvae of *Ascaris suum*. *Parasitol Res* **110,** 1583–1586 (2012).
27. Hewitson, J. P., Grainger, J. R. & Maizels, R. M. Helminth immunoregulation: the role of parasite secreted proteins in modulating host immunity. *Molecular and biochemical parasitology* **167,** 1–11, doi: 10.1016/j.molbiopara.2009.04.008 (2009).
28. Schwarz, E. M. *et al.* The genome and transcriptome of the zoonotic hookworm *Ancylostoma ceylanicum* identify infection-specific gene families. *Nat Genet* **47,** 416–422, doi: 10.1038/ng.3237 (2015).
29. Delannoy-Normand, A., Cortet, J., Cabaret, J. & Neveu, C. A suite of genes expressed during transition to parasitic lifestyle in the trichostrongylid nematode *Haemonchus contortus* encode potentially secreted proteins conserved in *Teladorsagia circumcincta*. *Vet Parasitol* **174,** 106–114 (2010).
30. Jing, Y., Toubarro, D., Hao, Y. & Simões, N. Cloning, characterisation and heterologous expression of an astacin metalloprotease, Sc-AST, from the entomoparasitic nematode *Steinernema carpocapsae*. *Molecular and Biochemical Parasitology* **174,** 101–108 (2010).
31. Delcroix, M. *et al.* A multienzyme network functions in intestinal protein digestion by a platyhelminth parasite. *Journal of Biological Chemistry* **281,** 39316–39329 (2006).
32. Geldhof, P., Claerebout, E., Knox, D., Agneessens, J. & Vercruysse, J. Proteinases released *in vitro* by the parasitic stages of the bovine abomasal nematode *Ostertagia ostertagi*. *Parasitology* **121,** 639–647 (2000).
33. Williamson, A. L. *et al.* Hookworm aspartic protease, Na-APR-2, cleaves human hemoglobin and serum proteins in a host-specific fashion. *Journal of Infectious Diseases* **187,** 484–494 (2003).
34. Aguileta, G., Refregier, G., Yockteng, R., Fournier, E. & Giraud, T. Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. *Infection, Genetics and Evolution* **9,** 656–670 (2009).
35. Rausell, A. & Telenti, A. Genomics of host-pathogen interactions. *Curr Opin Immunol* **30,** 32–38 (2014).

36. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution* **22,** 2472–2479 (2005).
37. Yang, Z., Wong, W. S. & Nielsen, R. Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular biology and evolution* **22,** 1107–1118 (2005).
38. Blaxter, M. L. *et al.* A molecular evolutionary framework for the phylum Nematoda. *Nature* **392,** 71–75 (1998).
39. Holterman, M. *et al.* Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown Clades. *Mol Biol Evol* **23,** 1792–1800 (2006).
40. Montiel, R., Lucena, M. A., Medeiros, J. & Simoes, N. The complete mitochondrial genome of the entomopathogenic nematode *Steinernema carpocapsae*: insights into nematode mitochondrial DNA evolution and phylogeny. *J Mol Evol* **62,** 211–225 (2006).
41. Nadler, S. A. *et al.* Phylogeny of Cephalobina (Nematoda): Molecular evidence for recurrent evolution of probolae and incongruence with traditional classifications. *Molecular Phylogenetics and Evolution* **40,** 696–711, doi: 10.1016/j.ympev.2006.04.005 (2006).
42. Hoffmann, M. *et al.* C. elegans ATAD-3 is essential for mitochondrial activity and development. *PLoS ONE* **4,** e7644 (2009).
43. Addo, M. G. *et al. Caenorhabditis elegans*, a pluricellular model organism to screen new genes involved in mitochondrial genome maintenance. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **1802,** 765–773 (2010).
44. Calderone, R., Li, D. & Traven, A. System-level impact of mitochondria on fungal virulence: to metabolism and beyond. *FEMS yeast research* **15,** fov027 (2015).
45. Bedding, R. Low cost *in vitro* mass production of *Neoaplectana* and *Heterorhabditis* species (Nematoda) for field control of insect pests. *Nematologica* **27,** 109–114 (1981).
46. Neves, J., Simoes, N. & Mota, M. Evidence for a sex pheromone in Steinemema carpocapsae. *Nematologica* **44,** 95–98 (1998).
47. Wright, S. Systems of mating. V. General considerations. *Genetics* **6,** 167 (1921).
48. Collins, G. G. & Symons, R. H. Extraction of nuclear DNA from grape vine leaves by a modified procedure. *Plant molecular biology reporter-ISPMB (USA)* (1992).
49. Sambrook, J., Fritsch, E. & Maniatis, T. *Molecular cloning: a laboratory manual*. (Cold Spring Harbor, 1989).
50. Bennett, M. D., Leitch, I. J., Price, H. J. & Johnston, J. S. Comparisons with Caenorhabditis (~100 Mb) and Drosophila (~175 Mb) using flow cytometry show genome size in Arabidopsis to be ~157 Mb and thus ~25% larger than the Arabidopsis genome initiative estimate of~125 Mb. *Annals of Botany* **91,** 547–557 (2003).
51. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0* http:/www.repeatmasker.org (2013).
52. Smit, A. F. A. & Hubley, R. *RepeatModeler Open-1.0*. http:/www.repeatmasker.org (2008).
53. Nawrocki, E. P. *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res* **43,** 11 (2015).
54. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29,** 2933–2935, doi: 10.1093/bioinformatics/btt509 (2013).
55. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25,** 0955–0964 (1997).
56. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35,** 3100–3108 (2007).
57. Guerra-Assunção, J. A. & Enright, A. J. MapMi: automated mapping of microRNA loci. *BMC bioinformatics* **11,** 133 (2010).
58. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research* **42(D41),** D68–D73 (2013).
59. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9,** e1003118 (2013).
60. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25,** 1841–1842 (2009).
61. Nguyen, K. & Hunt, D. *Entomopathogenic nematodes: systematics, phylogeny and bacterial symbionts*. (Brill, 2007).
62. Grabherr, M. G. *et al.* Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology* **29,** 644 (2011).
63. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31,** 5654–5666, doi: 10.1093/nar/gkg770 (2003).
64. Hao, Y.-J., Montiel, R., Abubucker, S., Mitreva, M. & Simões, N. Transcripts analysis of the entomopathogenic nematode *Steinernema carpocapsae* induced *in vitro* with insect haemolymph. *Molecular and Biochemical Parasitology* **169,** 79–86 (2010).
65. Davis, M. P., van Dongen, S., Abreu-Goodger, C., Bartonicek, N. & Enright, A. J. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods* **63,** 41–49 (2013).
66. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research* **41,** e108–e108 (2013).
67. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, gks042 (2012).
68. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11,** 1 (2010).
69. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Research* **19,** 1639–1645 (2009).
70. Tang, H. *et al.* SynFind: compiling syntenic regions across any set of genomes on demand. *Genome biology and evolution* **7,** 3286–3298 (2015).
71. Eric, L., Matthew, D. B., Shannon, L. O. & rew, J. L. In *Handbook of Plant and Crop Physiology*, Third Edition Books in Soils, Plants, and the Environment 797–816 (CRC Press, 2014).
72. Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research* **36,** 3420–3435, doi: 10.1093/nar/gkn176 (2008).
73. Camon, E. *et al.* The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Research* **13,** 662–672, doi: 10.1101/gr.461403 (2003).
74. Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *Journal of molecular biology* **340,** 783–795 (2004).
75. Kall, L., Krogh, A. & Sonnhammer, E. Advantages of combined 613 transmembrane topology and signal peptide prediction-the Phobius web server. *Nucleic Acids Res* **35,** W429–W432 (2007).
76. Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology* **305,** 567–580 (2001).
77. Emanuelsson, O., Nielsen, H., Brunak, S. & Von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of molecular biology* **300,** 1005–1016 (2000).
78. Horton, P. *et al.* WoLF PSORT: protein localization predictor. *Nucleic Acids Research* **35,** W585–W587 (2007).
79. Rawlings, N. D., Barrett, A. J. & Bateman, A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research* **40,** D343–D350 (2012).
80. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13,** 2178–2189 (2003).
81. Snyder, H., Stock, S. P., Kim, S.-K., Flores-Lara, Y. & Forst, S. New insights into the colonization and release processes of *Xenorhabdus nematophila* and the morphology and ultrastructure of the bacterial receptacle of its nematode host, *Steinernema carpocapsae*. *Applied and Environmental Microbiology* **73,** 5338–5346 (2007).

82. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical biochemistry* **72,** 248–254 (1976).
83. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32,** 1792–1797, doi: 10.1093/nar/gkh340 (2004).
84. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution* **17,** 540–552 (2000).
85. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27,** 1164–1165 (2011).
86. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19,** 1572–1574 (2003).
87. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
88. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* **52,** 696–704 (2003).
89. Wernersson, R. & Pedersen, A. G. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Research* **31,** 3537–3539 (2003).
90. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24,** 1586–1591 (2007).

Acknowledgements

This work was supported by a research grant from FOMIX-Hidalgo to RM (Fomix-Hgo-2008-C01-97032), and from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007/2013/ under REA grant agreement No. 612583. ARC received a postdoctoral fellowship from Conacyt (CVU 39220). MFP received a Conacyt fellowship for Master and Ph.D. studies (Reg. No. 219899). We are indebted to Sánchez-delPino M.M. and Valero L. from the proteomic service of the University of Valencia (Proteored ISCIII) for their support in proteomic data analysis. Clarisse, a script for massive scans of positive selection using PAML, was developed and kindly provided by Victor Villa-Moreno, from the "Laboratorio de la Diversidad Biomolecular" (Langebio) under the direction of Dr. Mauricio Carrillo-Tripp.

Author Contributions

A.R.C., M.F.P., R.M., designed the study; M.F.P., H.E.R.A., L.C., conducted experiments; A.R.C., M.F.P., H.E.R.A., C.E.M.G., Y.-J.H., J.A.R.M., C.O.V., J.R.B.B., C.A.G., R.M., analysed data; N.Ch.H., N.S., R.M., contributed materials and reagents; A.R.C., M.F.P., R.M., wrote the paper, with contributions from H.E.R.A., Y.-J.H., C.O.V., J.R.B.B., C.A.G. All authors reviewed and approved the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Rougon-Cardoso, A. *et al.* The genome, transcriptome, and proteome of the nematode *Steinernema carpocapsae*: evolutionary signatures of a pathogenic lifestyle. *Sci. Rep.* **6**, 37536; doi: 10.1038/srep37536 (2016).

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