A Lover and a Fighter: The Genome Sequence of an Entomopathogenic Nematode Heterorhabditis bacteriophora

Xiaodong Bai1*, Byron J. Adams2, Todd A. Ciche3, Sandra Clifton4,5†, Randy Gaugler6, Kwi-suk Kim3, John Spieth4,5, Paul W. Sternberg7, Richard K. Wilson4,5, Parwinder S. Grewal1*

1 Department of Entomology, The Ohio State University - OARDC, Wooster, Ohio, United States of America, 2 Department of Biology and Evolutionary Ecology Laboratories, Brigham Young University, Provo, Utah, United States of America, 3 Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan, United States of America, 4 Department of Genetics, Washington University School of Medicine, St Louis, Missouri, United States of America, 5 Genome Institute, Washington University School of Medicine, St Louis, Missouri, United States of America, 6 Department of Entomology, Rutgers University, New Brunswick, New Jersey, United States of America, 7 Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, Pasadena, California, United States of America

Abstract

Heterorhabditis bacteriophora are entomopathogenic nematodes that have evolved a mutualism with Photorhabdus luminescens bacteria to function as highly virulent insect pathogens. The nematode provides a safe harbor for intestinal symbionts in soil and delivers the symbiotic bacteria into the insect blood. The symbiont provides virulence and toxins, metabolites essential for nematode reproduction, and antibiotic preservation of the insect cadaver. Approximately half of the 21,250 putative protein coding genes identified in the 77 Mbp high quality draft H. bacteriophora genome sequence were novel proteins of unknown function lacking homologs in Caenorhabditis elegans or any other sequenced organisms. Similarly, 317 of the 603 predicted secreted proteins are novel with unknown function in addition to 19 putative peptidases, 9 peptidase inhibitors and 7 C-type lectins that may function in interactions with insect hosts or bacterial symbionts. The 134 proteins contained mariner transposase domains, of which there are none in C. elegans, suggesting an invasion and expansion of mariner transposons in H. bacteriophora. Fewer Kyoto Encyclopedia of Genes and Genomes Orthologies in almost all metabolic categories were detected in the genome compared with 9 other sequenced nematode genomes, which may reflect dependence on the symbiont or insect host for these functions. The H. bacteriophora genome sequence will greatly facilitate genetics, genomics and evolutionary studies to gain fundamental knowledge of nematode parasitism and mutualism. It also elevates the utility of H. bacteriophora as a bridge species between vertebrate parasitic nematodes and the C. elegans model.

Introduction

Nematodes are the most abundant multicellular animals on the planet [1], and exhibit remarkably diverse lifestyles to impact all life [2]. While some nematode parasites harm humans and agriculture, entomopathogenic (i.e., insect-parasitic) nematodes (EPNs) are beneficial in controlling insect pests [3,4]. Two EPN families, Heterorhabditidae and Steinernematidae, [5,6] have independently evolved mutual associations with insect pathogenic Photorhabdus and Xenorhabdus bacteria, respectively [7,8]. A specialized stage of the nematode, analogous to the C. elegans dauer, called the infective juvenile (IJ) harbors the mutualistic bacteria in its intestine while in search of an insect host [9]. Once found, the nematodes penetrate the insect body, sense unknown cue(s) in the hemolymph, and then regurgitate the symbionts [10,11]. The bacteria grow logarithmically and produce virulence factors and toxins causing rapid insect mortality [12–16]. The bacteria produce exoenzymes to degrade the insect tissues and produce unknown metabolites essential for nematode reproduction. Unlike C. elegans and other bacteria-feeding nematodes, H. bacteriophora reproduces only when associated with specific Photorhabdus bacteria both in insects and nutrient rich media [17,18]. In addition, the H. bacteriophora intestine is more permissive to symbiotic and non-symbiotic Escherichia coli OP50 intestinal bacteria than C. elegans [19]. The bacteria produce potent secondary metabolites that are antibiotics [20] and which deter
scaffolding arthropods [21], enabling the nematode proliferation to nearly 300,000 IJs from a single infected insect, which then disperse in search of new insect hosts [19,22].

**Heterorhabditis bacteriophora** and its mutualistic bacterium *Photorhabdus luminescens* represent a model system for the study of symbiosis and parasitism [11,23,24]. Although mutually dependent in nature, both organisms can be grown, manipulated and reassociated in culture. *Heterorhabditis* and *Photorhabdus* have congruent evolutionary lineages, indicating significant coevolution [25]. The bacteria adhere, persist, invade and grow inside nematode cells, breaching the alimentary tract to gain access to the developing IJs in the mother’s body [19]. The IJs select for bacteria that adhere to pharyngeal-intestinal valve cells, possibly invade these cells and exit to grow unattached in the intestinal lumen. It is likely that nematode receptors are exposed on specific cells in developmental stages where the bacteria adhere. For example, a phase variant subpopulation of the bacteria express maternal adhesion (Mad) fimbrae required for adhesion to the maternal intestine and transmission to IJs [26]. More surprisingly, the maternal nematodes select for a M-form phenotypic variant that is avirulent and slow growing compared to the insect pathogenic P form [27]. Visualizing the M-form cells persisting in the posterior intestine among the majority transients enabled the discovery that the P form changed to a small cell morphology (i.e. ~1/7 vol) of the M form. The optical transparency of the nematodes and differential labelling of transient and persistent bacteria made apparent the mutualistic function of phenotypic variant easily ignored. Furthermore, the genetic tractability of the symbiont and ease of screening revealed the mutable locus and transcription factors required for the P and M form switching [26]. It is unknown why nematodes acquire the M form, which switch genetically back to the P form in fully developed IJs and arm these nematodes for insect infection.

The IJs and bacteria endure cooperatively [27], often for many weeks to months without feeding [28] while in search for their host. Lowering their metabolism through cellular acidification and repressed motility may aid the bacteria to persist in the gut of the IJ [27]. In addition to vectoring the bacteria between insect hosts, the IJs may contribute to immune suppression of the insect hosts [29]. Thus, *H. bacteriophora* has evolved sophisticated adaptations for bacterial mutualism enabling it to function as an entomopathogen.

The availability of recent data on genome sequences has laid the necessary foundation for the development of this model system. The complete genome of *H. bacteriophora* strain TT01 symbiont, *Photorhabdus luminescens* subspp. laomondii TT01, was released in 2003 [30]. Transcriptomic data of *H. bacteriophora* TT01 and GPS11 recently became available [31–33]. Forward genetics by mutation of the promoter transcriptionally fused to beta-galactosidase [30] and DNA transformation [50] have been achieved. Transformation of the *H. bacteriophora* genome sequence reveals unique features that correspond to the evolution of mutualistic (lover) and parasitic (fighter) aspects of its biology.

### Results and Discussion

A total of 6,845,656 sequencing reads totaling 2,410,251,025 base pairs were obtained from *H. bacteriophora* genome. After quality trimming and assembly, a draft genome consisting of 1,263 scaffolds totaling 77,007,692 bp was obtained. The size of the scaffolds ranged from 327 to 2,228,510 bp with 166 scaffolds larger than 100 kb. The N50 value of the assembled genome is 312,328 bp. The overall GC content is 32.2%, which is similar to the free-living nematode *C. elegans*, plant-parasitic nematode *M. hapla*, and human-parasitic nematode *B. malayi* (Table 1). The assembly of *H. bacteriophora* genome sequence reveals unique features that correspond to the evolution of mutualistic (lover) and parasitic (fighter) aspects of its biology. *Heterorhabditis bacteriophora* is a transitional taxon among the Rhabditina. It exhibits ancestral traits shared with its microbivorous ancestors such as *C. elegans*, but has also evolved parasitism and shares most recent common ancestry with obligate mammalian parasites, such as hookworms and lungworms. Given this phylogenetic position, *Heterorhabditis* can serve as a sort of “bridge” taxon for exploring the evolutionary changes that free-living microbivores have undergone along the path to obligate parasitism of mammals (Figure 1A). Although this figure is not intended to be comprehensive, it does illustrate the general evolutionary trend from free-living microbivory through facultative and obligate associations with invertebrates, to obligate parasitism of vertebrates: *Panagrellus* represents a large clade of free-living microbivores, which gave rise to a series of subsequent evolutionary lineages that are non-parasitic associates of invertebrates, followed by *Heterorhabditis* and its sister taxon, the Strongylidea (represented by *Necator, Dictyocaulus* and *Olesurus*; obligate parasites of vertebrates). According to this scenario, a parsimonious reconstruction of evolutionary history features free-living microbivores giving rise to numerous microbivorous taxa that are facultative or opportunistic associates of invertebrates. However, such facultative and opportunistic conditions gave rise to a clade that evolved obligate parasitism. In *Heterorhabditis* microbivory (Figure 1B) and association with an invertebrate host were maintained. In contrast, the Strongylidea have lost microbivory during the evolution of obligate parasitism. However, the entomopathogenic symbiosis can also be viewed as an innovation in parasitism where nematode association with an insect pathogen increases the virulence and fitness of insect infection. The clade containing *Dictyocaulus* and *Olesurus* (lungworms; *Trichostrongylidae*, *Metastrongylidae*, respectively) has direct lifecycles, being ingested as larvae by their mammalian hosts [32–54]. *Necator* (Hookworms; *Ancylostomatidae*) penetrates tissue to infect its host. Most of the lungworms require an invertebrate (mollusk) intermediate host. Building on this foundation, the objective of this study was to obtain a high quality genome sequence to facilitate further insights into the mutualistic and parasitic lifestyles of *Heterorhabditis*. The analysis of *H. bacteriophora* genome sequence reveals unique features that correspond to the evolution of mutualistic (lover) and parasitic (fighter) aspects of its biology.

### Protein-coding Genes

The protein-coding genes were predicted using parameters optimized for *C. elegans* in the *ab initio* gene prediction programs. In total, 21,250 protein-coding genes were predicted (Table S1). The majority of the predicted protein genes, 11,207, had no significant homolog to *C. elegans* (WormBase release WS220), whilst 10,043 *H. bacteriophora* proteins had homologs with an E value cutoff of 1e-5 (Table S2). Of the protein-coding genes that have no homologs in WS220, 9,893 had no significant sequence similarity to known proteins in the GenBank non-redundant database and were hence considered novel. *H. bacteriophora* and strongyloid parasites like hookworms have adapted a developmentally arrested and alternative third larval stage, known as dauer larva in *C. elegans*, as the infective stage [53]. Entomopathogenic IJs harbor gut symbionts that benefit their
insect parasitism [56]. The *C. elegans* dauer develops under stressful conditions such as overcrowding by sensing dauer and other ascaroside pheromones, signal transduction through insulin and TGF-β pathways and DAF-12 nuclear hormone receptor [57–63]. *H. bacteriophora* produces an ascaroside ethanolamine (C11 EA) derivative that maintains the IJ state at high IJ densities and additional ascarosides [64,65]. We found that *H. bacteriophora* has most (19 of 23) of the insulin/IGF-1 signaling pathway genes that are critical for dauer formation and for regulation of longevity, stress resistance and innate immunity in *C. elegans* (Figure 2). We

---

**Table 1.** Comparison of *Heterorhabditis bacteriophora* genome with the complete genome of *Caenorhabditis elegans* (WS220) and the draft genomes of *Meloidogyne hapla* [132] and *Brugia malayi* [87].

<table>
<thead>
<tr>
<th>Life style</th>
<th>C. elegans</th>
<th>H. bacteriophora</th>
<th>M. hapla</th>
<th>B. malayi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome size, Mb</td>
<td>Free living</td>
<td>~ 80</td>
<td>Plant parasitic</td>
<td>Human parasitic</td>
</tr>
<tr>
<td>Scaffolds</td>
<td>n/a</td>
<td>1,263</td>
<td>54</td>
<td>90–95</td>
</tr>
<tr>
<td>Scaffold N50, bp</td>
<td>n/a</td>
<td>312,328</td>
<td>83,645</td>
<td>93,771</td>
</tr>
<tr>
<td>Assembled, bp</td>
<td>100,267,623</td>
<td>77,007,652</td>
<td>53,578,246</td>
<td>70,837,048</td>
</tr>
<tr>
<td>Gene models</td>
<td>21,193</td>
<td>21,250</td>
<td>14,420</td>
<td>11,515</td>
</tr>
<tr>
<td>Median exon, bp</td>
<td>147</td>
<td>112</td>
<td>145</td>
<td>140</td>
</tr>
<tr>
<td>Average exon/gene</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Median intron, bp</td>
<td>68</td>
<td>125</td>
<td>55</td>
<td>219</td>
</tr>
<tr>
<td>G+C, %</td>
<td>35.4</td>
<td>32.2</td>
<td>27.4</td>
<td>30.5</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0069618.t001
also found a daf-12 homolog predicted to function in ascaroside transcriptional response [66]. Study of IJ formation and exit from diapause, easily tested in insects like Drosophila melanogaster and assessed by release of intestinal symbionts [10,11], may lead to new antiparasitic strategies. Increasing IJ longevity and stress resistance may lead to improvements of EPNs for pest control [28,67,68].

RNA interference (RNAi) is a pathway for gene regulation and powerful tool to manipulate gene expression in functional genomics [69]. RNAi by soaking has been achieved in H. bacteriophora [24]. We detected sid-1 and sid-3 homologs required for systemic RNAi in C. elegans [70,71] but not a sid-2 homolog required in C. elegans for the uptake of dsRNA in the intestine [72]. Either an Hba- sid-2 homolog was left out of the current H. bacteriophora assembly or another transport mechanism is employed. Although C. elegans efficiently transports environmental DNA, most other related Caenorhabditis species do not [73]. Genes involved in RNA interference in H. bacteriophora, B. malayi, and M. hapla were identified based on sequence similarity to C. elegans gene products (Figure 3). Four genes, drsh-1, ego-1, rsd-3, and smg-2, have been identified in all four nematode species compared. In C. elegans, drsh-1 gene encodes a predicted RNase III-type ribonuclease that is orthologous to Drosha protein in Drosophila and human that is involved in processing primary miRNA transcripts (pri-miRNAs) in the nucleus [74]. ego-1 gene encodes putative RNA-directed RNA polymerase that is required for germline RNAi [75]. smg-2 is involved in non-sense-mediated mRNA decay that selectively and rapidly degrades eukaryotic mRNAs with premature stop codons [76]. rsd-3 is one of four RNA Spreading Defective genes (WormBase). A homolog of dcr-1 Dicer Related endonuclease [77] was detected in H. bacteriophora but not Dcr-1 associated protein rde-4, which is required for RNAi in C. elegans [78]. Since RNAi has been reported for H. bacteriophora [24], B. malayi [79,80], and M. hapla [81], different mechanisms are possibly employed.

**Protein Domains**

To begin to learn how the more than 10,000 unknown proteins function, we analyzed the proteins for conserved domains. A total of 7,957 Pfam domains with 4,144 different Pfam accessions were predicted using the program HMMER [82] with an E value cutoff of 1e-4. We compared the Pfam domains in H. bacteriophora with other nematodes [59] (Figure 4; Table S3). Based on protein domain information, we identified 82 members of GPCR (G protein coupled receptor) gene family and 24 members of NHR (nuclear hormone receptor) gene family. The domain richness index analysis (see methods) revealed 56 domains in H. bacteriophora that are significantly different from other nematodes. One significantly different richness domain index is the Mariner transposase (PF01359.11), with 138 identified in H. bacteriophora proteins compared to 65 in C. japonica, one each in M. incognita and M. hapla, but none in C. elegans and Brugia malayi. The Mariner transposases have been shown to be sufficient to mediate transposition in vitro in a purified form [83]. The enrichment of Mariner transposase domain is in agreement with the 1,314 predicted Mariner DNA motifs that belong to 23 types (Table 2;
symbiotic bacteria are required in the intestine for maternal
might also contain a diverse and novel set of innate immune
H. bacteriophora
pathogenic and saprophytic microorganisms.
P. luminescens
mutualistic relationship with
H. bacteriophora
domain proteins in
C. elegans
DNA motifs are exclusively present in
H. bacteriophora
genome (Table 2; Table S5). More strikingly, 28 types of Mariner
is likely evidence of a past or presently mobile
genomes.
We detected far fewer (9 vs. 133) C-type lectin domain-containing proteins than are present in C. elegans. Homologies of
lec-1, lec-2, lec-3, lec-5, lec-6, and lec-12 were detected that function in innate immunity in C. elegans [84]. The reduction in C-lectin
domain proteins in H. bacteriophora may be related to the mutualistic relationship with P. luminescens bacteria [19]. Viable
symbiotic bacteria are required in the intestine for maternal transmission and in IJs for insect infection. The H. bacteriophora
intestine is more permissive to symbiotic bacteria and non-symbiotic E. coli OP50 than C. elegans. Broad-spectrum antibiotics
produced by the symbionts likely contribute to defense against pathogenic and saprophytic microorganisms. H. bacteriophora
might also contain a diverse and novel set of innate immune effectors that were not detected by homology to C. elegans.

<table>
<thead>
<tr>
<th>Mariner type</th>
<th>Hba</th>
<th>Cel</th>
<th>Mariner type</th>
<th>Hba</th>
<th>Cel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mariner2_CE</td>
<td>36</td>
<td>93</td>
<td>Mariner36_CB</td>
<td>39</td>
<td>4</td>
</tr>
<tr>
<td>Mariner3_CE</td>
<td>18</td>
<td>73</td>
<td>Mariner37_CB</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Mariner4_CB</td>
<td>2</td>
<td>1</td>
<td>Mariner38_CB</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Mariner4_CE</td>
<td>27</td>
<td>8</td>
<td>Mariner38C_CB</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Mariner5_CE</td>
<td>4</td>
<td>68</td>
<td>Mariner38</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Mariner7_CB</td>
<td>–</td>
<td>180</td>
<td>Mariner40_CB</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Mariner8_CB</td>
<td>–</td>
<td>6</td>
<td>Mariner41_CB</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Mariner10_CB</td>
<td>–</td>
<td>3</td>
<td>Mariner42_CB</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Mariner12_CB</td>
<td>1</td>
<td>1</td>
<td>Mariner43_CB</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Mariner13_CB</td>
<td>59</td>
<td>9</td>
<td>Mariner44_CB</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Mariner14_CB</td>
<td>135</td>
<td>–</td>
<td>Mariner45_CB</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Mariner15_CB</td>
<td>332</td>
<td>1</td>
<td>Mariner47A_CB</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Mariner16_CB</td>
<td>108</td>
<td>–</td>
<td>Mariner47B_CB</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Mariner17_CB</td>
<td>12</td>
<td>–</td>
<td>Mariner47_CB</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Mariner18_CB</td>
<td>448</td>
<td>–</td>
<td>Mariner48_CB</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Mariner19_CB</td>
<td>12</td>
<td>–</td>
<td>Mariner51_CB</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Mariner20_CB</td>
<td>–</td>
<td>1</td>
<td>Mariner52_CB</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Mariner22_CB</td>
<td>2</td>
<td>4</td>
<td>Mariner53_CB</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>Mariner23_CB</td>
<td>–</td>
<td>1</td>
<td>Mariner54_CB</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Mariner25_CB</td>
<td>–</td>
<td>1</td>
<td>Mariner55_CB</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Mariner26_CB</td>
<td>–</td>
<td>1</td>
<td>Mariner56_CB</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Mariner27_CB</td>
<td>1</td>
<td>1</td>
<td>Mariner60_CB</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Mariner28_CB</td>
<td>–</td>
<td>166</td>
<td>Mariner61_CB</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Mariner31_CB</td>
<td>1</td>
<td>10</td>
<td>Mariner65_CB</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Mariner32_CB</td>
<td>64</td>
<td>1</td>
<td>Mariner66_CB</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Mariner34_CB</td>
<td>2</td>
<td>8</td>
<td>Total</td>
<td>1314</td>
<td>844</td>
</tr>
</tbody>
</table>

Abbreviations: Hba, Heterorhabditis bacteriophora; Cel, Caenorhabditis elegans.
doi:10.1371/journal.pone.0069618.t002

Non-coding RNA (ncRNA) and Regulatory Elements
A total of 134 potential microRNA (miRNA) genes were identified in H. bacteriophora genome representing 26 different animal microRNA species (Table S6). Other ncRNA include the U1, U2, U3, U4, U5, and U6 small nuclear RNA (snRNA) components of the spliceosome, SL1 involved in trans-splicing (none if 1e-10 cutoff is used), ribonuclease P (RNaseP), and eukaryotic-type signal recognition particle RNA. The number of the non-coding RNAs detected in H. bacteriophora is considerably less than those known to be present in C. elegans (Table S6). For instance, let-7 is absent in the current assembly although its presence and temporal expression were considered to be conserved among animals with bilateral symmetry [83], possibly due an incomplete genome assembly. The ncRNAs have important roles in regulating transcription, translation, and other biological processes.

A total of 254 transfer RNA (tRNA) genes and 1 tRNA pseudogene were predicted in H. bacteriophora genome by tRNAscan-SE (see Table S7) for all 20 standard amino acids, but not the tRNA-Selenocysteine gene. The number of detected tRNA genes in H. bacteriophora is dramatically lower than the 659 tRNA genes and at least 29 tRNA pseudogenes in C. elegans [86]. However, the number of tRNAs are close to those identified in human and plant parasitic nematodes. There are 233 tRNA genes and 26 tRNA pseudogenes identified in the human parasitic B. malayi [87] and 467 tRNA genes, 120 tRNA pseudogenes and 28 other tRNA genes in plant parasitic M. incognita [88].

Microsatellite Repeats
Microsatellites, also known as simple sequence repeats (SSRs), are tandem repeat sequences of 2–6 bp that serve as informative genetic markers to resolve relationships among closely related species because of their high mutation rate [89]. A total of 3,794 microsatellite loci were predicted in 506 contigs of the current draft H. bacteriophora genome (Table S8). Among them, 849 were located in coding regions. Previously, we developed 8 polymorphic microsatellite markers for H. bacteriophora that distinguished a Northeast Ohio population from other populations [90]. These microsatellite markers can serve as useful tools for determining the phylogeographic, demographic and genetic structure of H. bacteriophora populations.

Estimation of Divergence Time between H. bacteriophora and C. elegans
The divergence time between H. bacteriophora and C. elegans was estimated based on a set of 350 orthologs among H. bacteriophora, C. elegans, Anopheles gambiae, and Homo sapiens. Based on the divergence time of 800–1000 MYA between nematodes and insects [91], the estimated divergence time between H. bacteriophora and C. elegans is approximately 86–331 MYA. By contrast, the C. elegans and C. briggsae speciation date was estimated as 78–113 MYA [91]. The large (conservative) discrepancy between the upper and lower bounds are probably most strongly influenced by the sparse taxonomic sample (n = 4), as well as other analytical biases [92].

Characterization of the Secretome
H. bacteriophora secreted proteins are potentially important for parasitic interactions with insects, mutualistic interactions with symbiotic bacteria, immunity to pathogens and in development and reproduction. We detected 753 proteins with predicted signal peptides of which 150 also were predicted to be membrane localized. The 603 potentially secreted proteins (2.8% of total predicted proteins) are similar to the fraction of B. malayi secretome
proteins (2.3%), but are less than the free-living nematodes *C. elegans* (10.1%), *C. briggsae* (9.4%), *C. brenneri* (8.9%), *C. japonica* (6.2%), and *C. remanei* (8.8%), and the insect-associated *P. pacificus* (7.4%) when predicted with the same method and criteria. It is also about half of that of plant-parasitic nematodes *M. hapla* (5.2%) and *M. incognita* (5.2%). The low number of predicted secreted proteins in parasitic *H. bacteriophora* and *B. malayi* could be due to their reliance on mutualistic bacteria for these proteins.

Among the 603 *H. bacteriophora* secreted proteins, 164 had significant similarity (E value cutoff of 1e-5) to proteins in the SwissProt database (Table S9). Among the remaining 439 secreted proteins, 122 had significant similarity to proteins in the GenBank non-redundant database. The remaining 317 secreted proteins were novel proteins of unknown function. A search of the MEROPS database containing peptidases and peptidase inhibitors revealed the presence of 1 cysteine, 9 serine, and 9 metallo-peptidases and 9 peptidase inhibitors in *H. bacteriophora* secreted proteins (Table 3). Secreted peptidases have known roles in degrading host tissues for the benefit of parasites [92]. EPNs have been reported to release proteolytic enzymes to aid penetration of the insect gut to reach the hemocoel [93]. Following nematode penetration into the hemocoel, IJ secreted peptides and peptide inhibitors might function to disarm the insect serine proteinase cascade that results in pro-phenoloxidase activation and melanization, the elementary immune defense reaction [94]. However, during subsequent development of the nematode in the host hemocoel, the symbiont secretes peptidases/proteases [13–16,30], which may contribute to such functions. Indeed, the mutualistic bacteria of EPNs also act independently to suppress the insect immune system [29,95]. Therefore, both partners act synergistically in combating the insect immune system. A peptidase(s) also might function in utilizing symbiont-produced crystalline inclusion proteins (CipA and CipB) that are high in essential amino acid content and required for nematode reproduction [96]. *H. bacteriophora* also has homologs to *C. elegans* lysozyme genes *lys-1*, *lys-3–8* and *lys-10* that function in bacterial cell lysis and innate immunity [97]. Thus, although similarity suggests common function, it remains to be determined what roles most secreted proteins have in interspecies interactions.

Gene Ontology Enrichment

The predicted Gene Ontology of *H. bacteriophora* proteins was compared to those of the proteins from the other nine sequenced nematode genomes (Table S10). A striking difference is the significant enrichment of DNA metabolic process (GO:0006259), DNA recombination (GO:0006310), DNA-mediated transposition (GO:0006313), DNA integration (GO:0015074), transposition (GO:0032196) and transposase activity (GO:0004803) in *H. bacteriophora* compared to other nematodes, with the exception of *C. japonica*. These observations are in agreement with the enrichment of mariner transposase domain in *H. bacteriophora* discussed above.

Metabolic Pathway Comparison

The KEGG (Kyoto Encyclopedia of Genes and Genome) pathways were predicted for *H. bacteriophora* and other 9 nematode species for which full genome sequence information is available and the numbers of genes in each pathway are summarized in
The genes and KEGG orthology (KO) in the metabolic pathways were compared to assess whether there is enrichment or reduction in the *H. bacteriophora* genome compared to other select nematode genomes (Table 4). *H. bacteriophora* has fewer KOs compared to the free-living nematode *C. elegans* in almost all metabolic categories, which is compatible with previous observations that parasitic nematodes seem to undergo reductive genome evolution [98]. However, *H. bacteriophora* has substantially more proteins (48 in total) in the KO groups of glycan biosynthesis and metabolism (Table S12). Glycans are generally found attached to proteins as in glycoproteins and proteoglycans on the exterior surface of cells and play important roles in proper protein folding and cell-cell interactions [99]. At the enzyme level, *H. bacteriophora* has 17 (out of 23) enzymes in common with *C. elegans* (19 enzymes in total). Interestingly, *C. elegans, B. malayi* and *M. hapla* have only one isoform (isoform 1) of [heparan sulfate]-glucosamine 3-sulfotransferase (3-OST), whereas *H. bacteriophora* has three isoforms, isoform 1, 2 and 3. The enzyme 3-OST is involved in biosynthesis of glycan structure and different isoforms have been demonstrated to have different substrate specificities depending on the saccharide structures around the modified glucosamine residue [100]. The presence of the two additional isoforms of 3-OST enzyme together with other *H. bacteriophora*-specific enzymes involved in glycan biosynthesis and metabolism suggests that *H. bacteriophora* is well evolved to thrive in different environments where different metabolic substrates are available during its life cycle.

Orthologs

The orthologous sequences among *H. bacteriophora, C. elegans*, *C. briggsae*, *C. japonica*, *C. remanei*, *C. brenneri*, *B. malayi*, *Meloidogyne hapla*, *M. incognita*, and *Pristionchus pacificus* were identified using the orthoMCL program [101] on the predicted protein sequences from the genomes. In total, we identified 183 orthologs among these species (Table S13). Based on the Gene Ontology information of *C. elegans* genes in the ortholog sets, most of these orthologs are essential in *C. elegans*.

<table>
<thead>
<tr>
<th>Protein name</th>
<th>MEROPS family</th>
<th>Query start-end</th>
<th>MEROPS accession</th>
<th>Hit start-end</th>
<th>E value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hbpro09515</td>
<td>C46</td>
<td>185–256</td>
<td>MER011696</td>
<td>342–415</td>
<td>1.90e-07</td>
</tr>
<tr>
<td>Hbpro17338</td>
<td>M10A</td>
<td>122–237</td>
<td>MER003153</td>
<td>200–317</td>
<td>3.80e-30</td>
</tr>
<tr>
<td>Hbpro04992</td>
<td>M12A</td>
<td>92–170</td>
<td>MER003171</td>
<td>124–202</td>
<td>1.60e-20</td>
</tr>
<tr>
<td>Hbpro10653</td>
<td>M12A</td>
<td>94–207</td>
<td>MER015241</td>
<td>216–328</td>
<td>6.10e-25</td>
</tr>
<tr>
<td>Hbpro13863</td>
<td>M12A</td>
<td>319–441</td>
<td>MER024920</td>
<td>128–246</td>
<td>2.60e-23</td>
</tr>
<tr>
<td>Hbpro15592</td>
<td>M12A</td>
<td>133–320</td>
<td>MER002349</td>
<td>134–320</td>
<td>1.10e-09</td>
</tr>
<tr>
<td>Hbpro15986</td>
<td>M12A</td>
<td>145–326</td>
<td>MER001107</td>
<td>94–261</td>
<td>1.60e-37</td>
</tr>
<tr>
<td>Hbpro20263</td>
<td>M12A</td>
<td>70–253</td>
<td>MER001593</td>
<td>60–237</td>
<td>2.10e-41</td>
</tr>
<tr>
<td>Hbpro11857</td>
<td>M12B</td>
<td>104–164</td>
<td>MER002292</td>
<td>347–417</td>
<td>2.00e-25</td>
</tr>
<tr>
<td>Hbpro13918</td>
<td>M13</td>
<td>37–327</td>
<td>MER002350</td>
<td>78–370</td>
<td>1.60e-90</td>
</tr>
<tr>
<td>Hbpro01274</td>
<td>S01A</td>
<td>9–76</td>
<td>MER099499</td>
<td>26–91</td>
<td>2.80e-06</td>
</tr>
<tr>
<td>Hbpro17402</td>
<td>S08A</td>
<td>795–954</td>
<td>MER134526</td>
<td>298–451</td>
<td>8.30e-05</td>
</tr>
<tr>
<td>Hbpro16490</td>
<td>S08B</td>
<td>203–296</td>
<td>MER001610</td>
<td>179–272</td>
<td>1.20e-52</td>
</tr>
<tr>
<td>Hbpro11245</td>
<td>S09X</td>
<td>17–347</td>
<td>MER037861</td>
<td>26–353</td>
<td>1.20e-36</td>
</tr>
<tr>
<td>Hbpro11940</td>
<td>S09X</td>
<td>34–208</td>
<td>MER037861</td>
<td>29–207</td>
<td>4.70e-32</td>
</tr>
<tr>
<td>Hbpro20894</td>
<td>S10</td>
<td>35–61</td>
<td>MER000430</td>
<td>39–65</td>
<td>3.80e-07</td>
</tr>
<tr>
<td>Hbpro12626</td>
<td>S28</td>
<td>59–214</td>
<td>MER162965</td>
<td>54–211</td>
<td>3.50e-34</td>
</tr>
<tr>
<td>Hbpro14365</td>
<td>S28</td>
<td>130–242</td>
<td>MER171698</td>
<td>102–212</td>
<td>2.00e-40</td>
</tr>
<tr>
<td>Hbpro12626</td>
<td>S37</td>
<td>59–207</td>
<td>MER0001350</td>
<td>62–194</td>
<td>1.20e-05</td>
</tr>
</tbody>
</table>

Table S11. The genes and KEGG orthology (KO) in the metabolic pathways were compared to assess whether there is enrichment or reduction in the *H. bacteriophora* genome compared to other select nematode genomes (Table 4). *H. bacteriophora* has fewer KOs compared to the free-living nematode *C. elegans* in almost all metabolic categories, which is compatible with previous observations that parasitic nematodes seem to undergo reductive genome evolution [98]. However, *H. bacteriophora* has substantially more proteins (48 in total) in the KO groups of glycan biosynthesis and metabolism (Table S12). Glycans are generally found attached to proteins as in glycoproteins and proteoglycans on the exterior surface of cells and play important roles in proper protein folding and cell-cell interactions [99]. At the enzyme level, *H. bacteriophora* has 17 (out of 23) enzymes in common with *C. elegans* (19 enzymes in total). Interestingly, *C. elegans, B. malayi* and *M. hapla* have only one isoform (isoform 1) of [heparan sulfate]-glucosamine 3-sulfotransferase (3-OST), whereas *H. bacteriophora* has three isoforms, isoform 1, 2 and 3. The enzyme 3-OST is involved in biosynthesis of glycan structure and different isoforms have been demonstrated to have different substrate specificities depending on the saccharide structures around the modified glucosamine residue [100]. The presence of the two additional isoforms of 3-OST enzyme together with other *H. bacteriophora*-specific enzymes involved in glycan biosynthesis and metabolism suggests that *H. bacteriophora* is well evolved to thrive in different environments where different metabolic substrates are available during its life cycle.

Orthologs

The orthologous sequences among *H. bacteriophora, C. elegans*, *C. briggsae*, *C. japonica*, *C. remanei*, *C. brenneri*, *B. malayi*, *Meloidogyne hapla*, *M. incognita*, and *Pristionchus pacificus* were identified using the orthoMCL program [101] on the predicted protein sequences from the genomes. In total, we identified 183 orthologs among these species (Table S13). Based on the Gene Ontology information of *C. elegans* genes in the ortholog sets, most of these orthologs are essential in *C. elegans*. 

Table 3. Summary of secreted peptidases and peptidase inhibitors identified in *H. bacteriophora*.
and annotated to biological processes such as reproduction (number of orthologs: 50), growth (36), regulation of growth (47), regulation of biological process (61), and larval development (45). Genome sequences of other nematodes, including Bursaphelenchus xylophilus [102], Trichinella spiralis [90], and Ascaris suum [103], are not included in the analysis because trophic categories represented by these nematodes are already included in the current study.

**H. bacteriophora** is useful for Comparisons of Rapidly Evolving Protein Domains

Some proteins that are conserved from human to *C. elegans* have domains that are evolving too rapidly to analyze by the large evolutionary distance comparison. One example is the carboxyl terminal tail of EGF-receptor, called LET-23 in nematodes. A three-species comparison of *elegans-briggsae-japonica* has a C-terminus that is too conserved to be informative (being 65% identical), but addition of *H. bacteriophora* in a 4-way comparison highlights the tyrosines and PDZ-binding domain that have been shown to be functional in LET-23 [104,105], with only 26% identified across the four species (Figure 5).

**Conclusions**

*H. bacteriophora* is an entomopathogenic nematode, which is mutually associated with symbiotic bacteria to function as an insect parasite. The high quality draft genome sequence revolutionizes our knowledge and genetic tractability to understand nematode fundamental processes of gut mutualism and insect parasitism. *H. bacteriophora* is well-known of symbiosis compared to the *C. elegans* and thus represent a simple and tractable model of animal-bacteria gut symbiosis. The genome sequence along with RNAi gene silencing methodology provides a powerful reverse genetic approach to probe the functions of signaling pathways and transcription factors in symbiosis as well as insect parasitism. The *H. bacteriophora* genome sequence along with some sequences from other *H. bacteriophora* strains (e.g. GPS11) allow single nucleotide polymorphisms (SNPs) to be identified which can be used in mapping. For example, nematode mutations can be mapped to SNPs and identified by genome sequencing and their function validated by RNAi. In addition, *H. bacteriophora* cis- and untranslated regulatory elements can be identified and used to facilitate expression of transgenes. These approaches can be used to learn how the nematode associates with symbiotic bacteria, what is the basis for dependency of these nematodes on symbiotic bacteria for reproduction and how do nematodes function as parasites? Therefore, the *H. bacteriophora* TT01 genome facilitates both basic and applied research on entomopathogenic nematodes.

**Materials and Methods**

**Nematode Culture**

An inbred line, M31e, self-fertilized 13 times, of *H. bacteriophora* TT01 strain originally isolated from Trinidad and Tobago [106] and kindly provided by Dr. Ann Burnell (NUI-Maynooth, Ireland), was thawed from cryopreserved stocks [24]. Axenic IJs were obtained by culturing the nematodes on strain *P. temperata* TRN16 that do not colonize IJs [26]. High molecular weight DNA was purified from first and second larval stages harvested from lawns of TRN16 grown previously for 18 h at 28°C on NA-chol (4 g nutrient agar, 1 g sodium pyruvate, 10 g agarose per liter with 2 ml 5 mg/ml cholesterol added after autoclaving). On average, 275 IJs were added to 100 mm lawns for efficient egg laying. Nematodes were washed off the lawns after 82–86 h with 10 ml of Ringer’s containing 0.1% triton X-100. Bacteria were removed by

---

**Table 4. Genes and KEGG orthology (KO) in metabolic pathways in selected nematode species with different life styles.**

<table>
<thead>
<tr>
<th>Pathway</th>
<th>KOs in KEGG reference pathway</th>
<th>Genes</th>
<th>KOs</th>
<th>Genes</th>
<th>KOs</th>
<th>Genes</th>
<th>KOs</th>
<th>Genes</th>
<th>KOs</th>
<th>Genes</th>
<th>KOs</th>
<th>Genes</th>
<th>KOs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amino acid metabolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biosynthesis of other secondary metabolites</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carbohydrate metabolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Energy metabolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycan biosynthesis and metabolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lipid metabolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolism of other amino acids</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nucleotide metabolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xenobiotics biodegradation and metabolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**doi:10.1371/journal.pone.0069618.t004**
washing on a 10 μm pore nylon filter and hermaphrodites removed by retention on a 30 μm filter. Eggs were surface sterilized with 1% commercial bleach (Chlorox®), washed 3X in Ringer’s solution and allowed to hatch in Ringer’s solution containing 100 μg/ml carbenicillin, 50 μg/ml streptomycin, 30 μg/ml kanamycin and 10 μg/ml gentamicin overnight. A contaminant of Stenotrophomonas maltophilia, likely originating from a contaminated Ringer’s solution, was inadvertently sequenced along with the nematode. Approximately 3.610^6 L1 nematodes were harvested from 1,000 cultures.

Isolation of RNA

Nematode mRNA was isolated from mixed (L1–L4), adult and IJ stages grown on TRN16. The nematodes were obtained from the cultures with Ringer’s solution and bacteria removed by 3 washes with 15Ringer’s solution in a 15 ml conical tube and centrifugation for 5 min at 2,000 rpm. The nematodes were frozen in liquid nitrogen, then Trizol reagent (Life Technologies) was added and incubated at 65°C. The IJs were freeze-thawed 3X in liquid nitrogen and at 65°C, before RNA was purified per manufacturer’s instructions. Polyadenylated RNA was purified using oligo(dT) cellulose columns, MicroPoly(A)Purist Kit (Life Technologies).

cDNA Library Construction and Sequencing

The integrity of the mRNA was validated using the Bioanalyzer 2100 (Agilent Technologies) and yield determined via Nanodrop (Thermo Scientific). Two different methods were used for library construction:

1) The CloneMiner cDNA Library Construction Kit (Life Technologies) was utilized to generate non-radiolabeled cDNA according to the manufacturer’s specifications. A Biotin-attB2-Oligo(dT) primer was hybridized to mRNA. First strand cDNA was synthesized via SuperScript II Reverse Transcriptase. DNA polymerase I was utilized to generate the second strand of cDNA. attB1 adapters were ligated to the 5’ end of the cDNA. The cDNA was purified by column fractionation to remove residual adapters. Through site-specific recombination, attB-flanked cDNA was cloned directly into the pDONR-222 vector (Life Technologies). The ligations were transformed using the ElectroMax DH10B cells (Life Technologies). The transformed cells were spread on LB plates containing 50 μg/mL kanamycin.

2) mRNA was used as the template for cDNA library construction using the Accuscript HF Reverse Transcriptase Kit (Agilent Technologies) and SMART primers (Life Technologies). PCR cycle optimization was performed to determine the threshold cycle number to minimally amplify full length cDNA products using the SMART primers and Clontech Advantage-HF 2 polymerase Mix (Clontech/Takara Bio). Library normalization was accomplished by using the Trimmer kit (Evrogen). PCR cycle optimization was performed with normalized cDNA to determine the threshold cycle number using the SMART primers and Clontech Advantage-HF 2 polymerase Mix previously mentioned. Finally, 5’ and 3’ adapter excision was performed by restriction exonuclease digestion using MmeI. The excised adapters were removed utilizing AMPure paramagnetic beads (Agencourt, Beckman Coulter Genomics). Two kinds of libraries were prepared for sequencing on ABI3730 and Roche/454 platforms.

Figure 5. H. bacteriophora informs C. elegans protein structure function. Multiple alignment of the EGF-receptor (LET-23) carboxyl tail of Caenorhabditis elegans, briggsae and japonica with H. bacteriophora. 3-way, alignment of the three Caenorhabditis proteins; 4-way, alignment of three Caenorhabditis proteins with Hba-LET-23. *, identity; : strong similarity; ., weak similarity. Red and green highlight the parts of the protein that have been demonstrated to be important in signaling and localization, respectively. Numbers represent the length of the predicted proteins.
doi:10.1371/journal.pone.0069618.g005
For libraries intended for sequencing on ABI3730 platform, the final cDNA product was nebulized, end repaired (Lucigen), and size selected from a 0.8% SeaKem agarose TAE gel. The fraction was purified according to the manufacturer's instructions in the QiAquick Gel Extraction (Qiagen) protocol and ligated into the pSMART HC-Kan vector system (Lucigen). Ligation was transformed using E. coli cells (Lucigen). The transformed cells were spread onto LB plates containing 50 μg/mL kanamycin.

A 454 fragment library was constructed using GS DNA Library Preparation Kit (Roche) with the cDNA as outlined in the manufacturer's protocol. Five microgram of cDNA was fragmentated via nebulization. Fragmented cDNA was size selected with an AMPure bead (Agencourt, Beckman Coulter Genomics) cleanup, removing fragments less than 300 bp. The cDNA was end polished and ligated to 454 Titanium library adapters utilizing reagents from the Titanium General Library Kit (Roche). An AMPure (Agencourt) bead cleanup was performed to remove library adapter dimers and cDNA fragments less than 400 bp in length. The 454 library was immobilized with Streptavidin beads (- Roche) and single stranded with Sodium Hydroxide. The single stranded library was quantitated by a Quant-iT single stranded DNA assay using a Qubit fluorometer (Life Technologies) and the integrity validated using the BioAnalyzer 2100 (Agilent Technologies). The library fragments were immobilized onto DNA capture beads utilizing clonal amplification kits (Roche). The captured DNA library was emulsified and subjected to PCR in order to amplify the DNA template. The emulsion was chemically broken and the beads containing the DNA were recovered, washed, and enriched utilizing bead recovery reagents (Roche). The DNA library beads were loaded onto a PicoTiterPlate device and sequenced on the Genome Sequencer instrument using the GS FLX Titanium Sequencing Kit XLR70 (Roche).

Genomic Library Construction and Sequencing

High molecular weight genomic DNA was isolated using a protocol kindly provided by Erich Schwartz, which was based on that of Andrew Fire's lab with slight modifications from the R. Waterston lab and K. Kiontke [107]. The integrity of the genomic DNA was verified by comparing the intensity of H. bacteriophora to serial dilutions of lambda standards of known concentration on a 1.8% agarose gel stained with ethidium bromide. The yield was determined by a high sensitivity Quant-iT single stranded DNA assay using the Qubit fluorometer (Life Technologies) and the integrity validated using the BioAnalyzer 2100 (Agilent Technologies). The library fragments were immobilized onto DNA capture beads utilizing clonal amplification kits (Roche). The captured DNA library was emulsified and subjected to PCR in order to amplify the DNA template. The emulsion was chemically broken and the beads containing the DNA were recovered, washed, and enriched utilizing bead recovery reagents (Roche). The DNA library beads were loaded onto a PicoTiterPlate device and sequenced on the Genome Sequencer instrument using the GS FLX Titanium Sequencing Kit XLR70 (Roche).

Genome Assembly

The genome sequences from fragments, 3 kb insert from plasmid libraries and end sequencing of bacterial artificial clone libraries were generated at an estimated 26-fold sequence coverage. All sequenced reads were attempted in de novo assembly using the Celera assembler v. 6.0. The assembly was submitted to GenBank genome database under accession number ACKM00000000.

Genome Annotation

The scaffolds were masked for repeats using RepeatMasker version 3.3 [108]. Transfer RNA coding genes were predicted using tRNAscan-SE [109]. To identify microRNA, other non-coding RNA, and regulatory elements, Rfam [110] covariance models were searched using Inferno program [111,112] with an E value cutoff of 1e-8 after adjusting to the size of the genome. Protein-coding genes were predicted with gene prediction programs of SNAP [113], AUGUSTUS [114–116], GlimmerHMM [117], and GeneMark [118]. The results were integrated with other evidence, including the mapping results of ESTs generated by cDNA sequencing with simil sequence similarity to proteins in GenBank non-redundant (nr) database and WormBase WS220 release, by JIGSAW program [119] with linear combiner option. Gene models with in-frame stop codons were considered erroneous and therefore removed. Protein domains in the predicted protein-coding genes were predicted by searching Pfam [120] using the HMMER program [82] with an E value threshold of 1e-4. For comparison, the same prediction parameters were used to predict Pfam domains in other nematodes. A domain richness index for each domain in each nematode was calculated by dividing the number of that domain with the total number of protein sequences in that nematode species. The program T statistics was used to compare the domain richness indices among nematodes. H. bacteriophora protein sequences were assigned Gene Ontology terms by the Blast2GO program [121] based on the BLASTp results against the SwissProt database with an E value cutoff of 1e-10. The orthologous sequences among H. bacteriophora, C. elegans, C. briggsae, C. jenensis, C. remanei, Brugia malayi, Meloidogyne hapla, M. incognita, and Pristionchus pacifica were identified using the orthoMCL program [101] on the predicted protein sequences from the genomes. H. bacteriophora protease/peptidases were predicted based on sequence similarity search of the sequences in MEROPS database Release 9.5 [122].

Estimation of Divergence Time between H. bacteriophora and C. elegans

We obtained a set of 350 orthologs common to H. bacteriophora, C. elegans, Apotheles gambiae (AgamP3.4 release from VectorBase), and Homo sapiens (Ensembl release 55) based on the prediction results of orthoMCL [101]. For each ortholog set, the protein sequences were aligned using ChustaW2 [123], followed by reverse translation to their original transcript sequences that were obtained from the same respective databases as the protein sequences. After conversion to PHYLIP format, the alignments were used to estimate genetic distances among the taxa using the DNADIST program in PHYLIP (PHYLogeny Inference Package; [124]). A phylogenetic tree was then built using the PHYLIP neighbor-joining algorithm NEIGHBOR with human as the
outgroup taxon. The sequence alignment and the rooted neighbor-joining tree were used to estimate divergence times using the MCMC TREE program in PAML (Phylogenetic Analysis by Maximum Likelihood [125]). We used 900–1000 MYA (million years ago) as the divergence time of nematodes and insects [91].

Gene Ontology Enrichment and Metabolic Pathway Comparison

*H. bacteriophora* protein sequences were assigned Gene Ontology (GO) terms by the Blast2GO program [121] based on the BLASTP results against SwissProt database with an E value cutoff of 1e-10. In comparison, proteins from the other 9 nematode genomes underwent the same analysis using the same programs and databases. The pair-wise GO enrichment using *H. bacteriophora* sequences as the reference was done using the GOSSIP program [126]. The KEGG (Kyoto Encyclopedia of Genes and Genomes) Ontologies (KO) in the metabolic pathways were assigned using Blast2GO program [121] for the four nematode species being compared.

Ethics Statement

This study did not involve any human or vertebrate subjects.

Supporting Information

Table S1 Predicted gene models in *Heterorhabditis bacteriophora* genome.

Table S2 Sequence similarity of conceptually translated *H. bacteriophora* proteins to *C. elegans* proteins in Wormbase release W220.

Table S3 Comparison of Pfam domains predicted in the proteins of 10 nematode species in this study.

Table S4 Predicted mariner DNA motifs in *H. bacteriophora* genome.

Table S5 Predicted mariner DNA motifs in *C. elegans* genome using the same parameters as the ones used to generate results in Table S4.

References


100. Liu J, Shworak NW, Sinay P, Schwartz JJ, Zhang L, et al. (1999) Expression of...
93. AbuHatab M, Selvan S, Gaugler R (1995) Role of Proteases in Penetration of...
91. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, et al. (2003) The...
81. Huang G, Allen R, Davis EL, Baum TJ, Hussey RS (2006) Engineering broad...
80. Song C, Gallup JM, Day TA, Bartholomay LC, Kimber MJ (2010)...
75. Vought VE, Ohmachi M, Lee MH, Maine EM (2005) EGO-1, a putative...
96. Bintrim SB, Ensign JC (1998) Insertional inactivation of genes encoding the...
235.
129. Kiontke K, Fitch DHA (2005) The phylogenetic relationships of Caenorhabditis...
128. Elsworth B, Wasmuth J, Blaxter M (2011) NEMBASE4: the nematode...
119. Allen JE, Salzberg SL (2005) JIGSAW: integration of multiple sources of...
118. Allen JE, Salzberg SL (2005) JIGSAW: integration of multiple sources of...