Published in final edited form as:

Nat Genet. 2016 March ; 48(3): 299–307. doi:10.1038/ng.3495.

The Genomic Basis of Parasitism in the *Strongyloides* Clade of Nematodes

Vicky L. Hunt[1,a], **Isheng J. Tsai**[2,3,a], **Avril Coghlan**[4,a], **Adam J. Reid**[4,a], **Nancy Holroyd**[4], **Bernardo J. Foth**[4], **Alan Tracey**[4], **James A. Cotton**[4], **Eleanor J. Stanley**[4], **Helen Beasley**[4], **Hayley M. Bennett**[4], **Karen Brooks**[4], **Bhavana Harsha**[4], **Rei Kajitani**[5], **Arpita Kulkarni**[6], **Dorothee Harbecke**[6], **Eiji Nagayasu**[3], **Sarah Nichol**[4], **Yoshitoshi Ogura**[7], **Michael A. Quail**[4], **Nadine Randle**[8], **Dong Xia**[8], **Norbert W. Brattig**[9], **Hanns Soblik**[9], **Diogo M. Ribeiro**[4], **Alejandro Sanchez-Flores**[4,10], **Tetsuya Hayashi**[7], **Takehiko Itoh**[5], **Dee R. Denver**[11], **Warwick Grant**[12], **Jonathan D. Stoltzfus**[13], **James B. Lok**[13], **Haruhiko Murayama**[3], **Jonathan Wastling**[8,14], **Adrian Streit**[6], **Taisei Kikuchi**[3], **Mark Viney**[1], and **Matthew Berriman**[4]

[1]School of Biological Sciences, University of Bristol, Bristol, BS8 1TQ, UK.

[2]Biodiversity Research Center, Academia Sinica, Taipei 11529, Taiwan.

[3]Division of Parasitology, Faculty of Medicine, University of Miyazaki, Miyazaki, Japan.

Corresponding authors: Taisei Kikuchi, Taisei_kikuchi@med.miyazaki-u.ac.jp; Mark Viney, Mark.Viney@bristol.ac.uk; Matthew Berriman mb4@sanger.ac.uk.
[a]Equal contributors

URLs

WormBase ParaSite, http://parasite.wormbase.org/.
WHO Soil-transmitted helminthiases, http://www.who.int/gho/neglected_diseases/soil_transmitted_helminthiases/en/
WHO | Estimates for 2000–2012, http://www.who.int/healthinfo/global_burden_disease/estimates/en/index2.html

Accession Codes

The *S. ratti, S. stercoralis, S. papillosus, S. venezuelensis, P. trichosuri* and *Rhabditophanes* genome assemblies, predicted transcripts, protein and annotation (*.GFF) files are available from WormBase ParaSite and are registered under BioProject accessions PRJEB125 (S_ratti_ED321_v5_0_4), PRJEB528 (S_stercoralis_PV0001_v2_0_4), PRJEB525 (S_papillosus_LIN_v2_1_4), PRJEB530 (S_venezuelensis_HH1_v2_0_4), PRJEB515 (P_trichosuri_KNP_v2_0_4) and PRJEB1297 (Rhabditophanes_sp_KR3021_v2_0_4). The raw genomic data are available from the ENA via accession numbers detailed in Supplementary Table 23.
The transcriptomic data are available from ArrayExpress under accession numbers E-ERAD-151 and E-ERAD-92 (*S. ratti*) and the DRA under accession number PRJDB3457 (*S. venezuelensis*) (Supplementary Table 24).

Author contributions

Cultivated and collected parasite material: V.L.H., D.D., W.G., J.B.L., E.N., H.M., A.S., A.K., J.D.S., D.H., T.K., M.V. Prepared DNA, RNA and protein: V.L.H., N.R., T.K., Prepared libraries: M.A.Q., H.B., E.N., Y.O., J.D.S. Assembled the genomes: I.J.T., A.S.F., R.K., T.I. Quality-checked the genomes: A.C. Provided genetic markers and mapping data: A.K., D.H., A.S. Manually improved the genomes: A.T., H.B., K.B., S.N., I.J.T. Predicted the genes: E.S., A.C., B.J.F., I.J.T. Functionally annotated the genome: A.C., D.R., B.H. Curated gene models: A.T., H.B., K.B., S.N., I.J.T., V.L.H. Built a Compara database: B.H. Analyzed gene structure: I.J.T. Undertook proteomics and initial analysis: N.R., D.X., N.B., J.W., V.L.H. Undertook excretory/secretome work and analyzed the data: N.W.B., H.S., D.X., V.L.H. Analyzed the transcriptome: V.L.H., I.J.T., B.J.F., A.J.R., J.D.S. Analyzed the gene clusters: D.M.R., V.L.H. Analyzed synteny and chromosome alignments: I.J.T., A.J.R. Assembled and analyzed mitochondrial genomes: T.K., I.J.T. Chromatin diminution analysis: B.J.F., I.J.T., A.C. Analyzed gene family clustering: A.J.R., J.A.C., I.J.T. Coordinated the project, managed sequencing, assembly and finishing: N.H., T.H., T.K. Wrote the manuscript: V.L.H., I.J.T., A.C., A.J.R., N.H., T.K., M.V., M.B. Conceived the project: M.V., M.B., J.W., I.J.T., T.K. Directed the project: M.V., M.B.

URLs

RepeatModeler, http://www.repeatmasker.org/RepeatModeler.html/;
TransposonPSI, http://transposonpsi.sourceforge.net/.
SMALT, www.sanger.ac.uk/resources/software/smalt/.

Competing Financial Interests

The authors declare no competing financial interests.

[4]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, UK.

[5]Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan.

[6]Max Planck Institute for Developmental Biology, Tübingen, Germany.

[7]Department of Bacteriology, Faculty of Medical Sciences, Kyushu University, Fukuoka, Japan.

[8]Department of Infection Biology, Institute of Infection and Global Health and School of Veterinary Science, University of Liverpool, Liverpool, UK.

[9]Department of Molecular Medicine, Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany.

[10]Unidad de Secuenciación Masiva y Bioinformática, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México, 62210.

[11]Department of Intergrative Biology, Oregon State University, Corvallis, Oregon, USA.

[12]Department of Animal, Plant and Soil Sciences, La Trobe University, Melbourne, Victoria, Australia.

[13]Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, 3800 Spruce Street, Philadelphia 19104, PA, USA.

[14]Faculty of Natural Sciences, University of Keele, Keele, Staffordshire, ST5 5BG, UK.

Abstract

Soil transmitted nematodes, including *Strongyloides*, cause one of the most prevalent Neglected Tropical Diseases. Here we compare the genomes of four *Strongyloides* spp., including the human pathogen *S. stercoralis*, and their close relatives that are facultatively parasitic (*Parastrongyloides trichosuri*) and free-living (*Rhabditophanes* sp). A significant paralogous expansion of key gene families – astacin-like and SCP/TAPS coding gene families – is associated with the evolution of parasitism in this clade. Exploiting the unique *Strongyloides* life cycle we compare the transcriptome of its parasitic and free-living stages and find that these same genes are upregulated in the parasitic stages, underscoring their role in nematode parasitism.

Keywords

Helminth; nematode; genome; *Strongyloides*; *Parastrongyloides*; *Rhabditophanes*; parasitism; transcriptome; proteome; synteny; astacins; SCP/TAPS; gene clusters

Introduction

More than a billion people are infected with intestinal nematodes[1,2]. The World Health Organization (WHO) has classified infections with soil transmitted nematodes[3] as one of the 17 most neglected tropical diseases and estimates that worldwide they cause an annual disease burden of 5 million Years Lost due to Disability (YLD), greater than that for malaria

(4 million YLD) and HIV/AIDS (4.5 million)[4]. Parasitic nematode infections can impair physical and educational development[1].

Strongyloides spp. are soil-transmitted gastrointestinal parasitic nematodes infecting a wide range of vertebrates[5]. Two species – *S. stercoralis* and *S. fuelleborni* – infect some 100–200 million people worldwide[6,7]. Other *Strongyloides* species infect livestock, such as *S. papillosus* infection in sheep.

Strongyloides spp. are from a clade of nematodes[8–10] that include taxa with diverse lifestyles including free-living (*Rhabditophanes*), parasitism of invertebrates, facultative parasitism of vertebrates (*Parastrongyloides*) and obligate parasitism of vertebrates (*Strongyloides*)[8,9]. Nematodes have independently evolved parasitism of animals several times[11], and thus understanding the genomic adaptations to parasitism in one clade will help in understanding how parasitism has evolved across the phylum more widely.

The *Strongyloides* life cycle alternates between free-living and parasitic generations. The female only, parthenogenetic[12] parasitic stage lives in the small intestine of its host where it produces offspring that develop outside of the host either directly to infective third-stage larvae (iL3s) or via a dioecious, sexually reproducing, adult generation[13], whose progeny are also iL3s. The iL3s penetrate the skin of a host and migrate to its gut[14] where they develop into parasitic adults (Fig. 1). Therefore, this life cycle has two genetically identical adult female stages – one obligate and parasitic, and one facultative and free-living; we have compared these transcriptomically and proteomically to reveal the genes and gene products specifically present in the parasitic stage. The closely related genus *Parastrongyloides*[5,15] is similar to *Strongyloides* spp., except that its parasitic generation is dioecious and sexually reproducing, and that it can have apparently unlimited cycles of its free-living adult generation[5,16](Fig. 1).

Here we report the genome sequences for six nematodes from one superfamily: four species of *Strongyloides* – *S. stercoralis* (a parasite of humans and dogs), *S. ratti* and *S. venezuelensis* (both parasites of rats, and important laboratory models of nematode infection) and *S. papillosus* (a parasite of sheep); *Parastrongyloides trichosuri* (which infects the brushtail possum *Trichosurus vulpecula*), and the free-living nematode *Rhabditophanes* sp.[8].

To investigate the genomic and molecular basis of parasitism in these nematodes we compared (i) the genomes and gene families of these parasitic (*Strongyloides* and *Parastrongyloides*, the Strongyloididae) and free-living (*Rhabditophanes*) taxa (Fig. 1); (ii) the transcriptomes of parasitic adult females, free-living adult females and iL3s of *S. ratti* and *S. stercoralis*, and (iii) the proteomes of parasitic and free-living females of *S. ratti*. We have identified the genes present in the parasitic species, and the genes and gene products uniquely upregulated in the parasitic stages of *S. stercoralis* and *S. ratti*; together these are the major genomic and molecular adaptations to the parasitic lifestyle of these nematodes.

Results

Chromosome biology

We have produced a high-quality 43 Mb reference genome assembly for *S. ratti* (Supplementary Note), with its two autosomes[17] assembled into single scaffolds and the X chromosome[17] into ten (Table 1; Fig. 2). This assembly is the second most contiguous assembled nematode genome after the *Caenorhabditis elegans* reference genome[18]. We also produced high quality draft assemblies of the 42–60 Mb genomes of *S. stercoralis, S. venezuelensis, S. papillosus, P. trichosuri* and *Rhabditophanes* sp., which are 95.6 – 99.6% complete (Supplementary Table 1). With GC contents of 21% and 22% respectively, the *S. ratti* and *S. stercoralis* genomes are the most AT-rich reported to date for nematodes (Supplementary Table 1). The ~43 Mb *S. ratti* and *S. stercoralis* genomes are small compared with other nematodes. However, the total protein-coding content of each nematode genome is similar (18–22 Mb versus 14–30 Mb in eight outgroup species; Supplementary Table 1). Significant loss of introns as well as shorter intergenic regions account for the smaller genomes from the present study (Spearman's correlation between genome size and intron number $\rho=0.91$, $P < 0.001$ and size of intergenic regions $\rho=0.63$, $P = 0.02$; Supplementary Table 2). However, parsimony analysis of intronic positions conserved in two or more species revealed that substantial intron losses occurred prior to the evolution of the *Rhabditophanes-Parastrongyloides-Strongyloides* clade (Supplementary Fig. 1), and are therefore not an adaptation associated with parasitism.

The canonical view of a nematode chromosome, defined nearly twenty years ago using *C. elegans* autosomes (and later confirmed in *C. briggsae*[19]) is of a gene-dense, repeat-poor "center" of conserved genes (based on homology with yeast genes[18]), flanked by two gene-poor, repeat-rich "arms" in which most genes are less strongly conserved. *S. ratti* is the first non-*Caenorhabditis* nematode whose whole chromosomes have been assembled and it presents a strikingly different organisation with relatively little variability in gene density, repeat density or gene conservation to yeast genes along its autosomes (Supplementary Figs. 2, 3).

Synteny is highly conserved within the parasitic Strongyloididae, but much less between this family and *Rhabditophanes* (Fig. 2). Scaffolds of the parasitic species largely correspond to blocks from a particular *S. ratti* chromosome, but in a scrambled order. This suggests that intra-chromosomal rearrangement is frequent, but inter-chromosomal rearrangement is rare, a common phenomenon in nematode chromosome evolution[19–21]. The notable exception was for *S. papillosus* and *S. venezuelensis* scaffolds that have many blocks that are syntenic to both *S. ratti* chromosome I and X (Supplementary Table 3). This likely reflects the fusion event between chromosomes I and X in these species[22–24]. Associated with this fusion is a change in the chromosome biology of sex determination in these species. *S. papillosus* undergoes chromatin diminution (where a chromosome fragments after which part of the chromosome is eliminated during mitosis) to mimic the XX/XO sex-determining system of *S. ratti*[25] and *S. stercoralis*[22].

By analyzing the differential coverage of mapped sequence data from iL3s (which are all female) and adult males, we were able to identify regions of the *S. papillosus* X-I fusion

chromosome that are eliminated from males during diminution (Supplementary Table 4). Six scaffolds were identified from the diminished region using existing genetic markers (Supplementary Table 5), but our read-depth approach extended this map to 153 scaffolds (18% of the assembly, 10.9 Mb). Interestingly, some genes with orthologs on the X chromosome of *S. ratti* are not diminished in *S. papillosus*, so dosage of these genes in males has changed since the species diverged, including three genes on *S. papillosus* chromosome II (confirming earlier work[22]), and 33 that lie in non-diminished regions of the X-I fusion chromosome (Supplementary Table 6).

Extensive rearrangement of the mitochondrial gene order

The *S. stercoralis* mitochondrial (mt) genome is highly rearranged compared with nematodes from clades I, III and V[26]. Manual finishing of the mt genomes of the six species revealed that the *Rhabditophanes* mt genome consists of two circular chromosomes, a feature of some other nematode species[27]. Compared with eight outgroup species, *Rhabditophanes* has a conventional gene order but *Strongyloides* spp. and *P. trichosuri* have highly rearranged mt genomes (Fig. 2, Supplementary Table 7). Similar observations have been reported in other clade IV parasitic nematodes[27–30] and there is evidence of mt recombination[29,31], which is rarely observed in animals[32]. Consistent with published nematode mt genomes, the gene-based phylogeny of the mt genome (Fig. 2) conflicts with phylogenies based on nuclear genes[29,33,34], and the rearranged gene order of the mt genome of *Strongyloides* spp. is accompanied by nucleotide divergence (Fig. 2).

Gene families associated with the evolution of parasitism

We predicted 12,451–18,457 genes across the six genomes, numbers comparable to other nematode species (Table 1, Supplementary Fig. 4). We then used Ensembl Compara (Supplementary Note)[35] to identify orthologs and gene families (Supplementary Table 8) in these and eight outgroup species, encompassing four further nematode clades (Supplementary Fig. 4). By pinpointing when a new gene family arose, and where a family has expanded or contracted, we could determine which gene families are associated with the evolution of parasitism. The largest acquisition of gene families (1075 families) was found on the branch leading to the parasitic nematodes, *Strongyloides* spp. and *P. trichosuri* (Fig. 1, Supplementary Fig. 4). Despite this highly dynamic pattern of gene gains and loss within each species' genome, the proportion of *Strongyloides*- (and Strongyloididae-) specific genes is consistent across the phylogeny (Fig. 1). The branches leading to these five parasitic species also showed greater expansion of genes and families of genes, compared to that in the free-living *Rhabditophanes*. Gain and expansion of gene families in these parasitic species likely reflects the necessary adaptations required by these species to be able to parasitize vertebrate hosts while maintaining a free-living phase.

The two most expanded *Strongyloides* spp. gene families encode astacin-like[36] and SCP/TAPS (SCP/Tpx-l/Ag5/PR-1/Sc7[37], also known as CAP-domain) proteins, present in multiple subfamilies (based on Ensembl Compara analysis, Supplementary Table 8, and protein domain combinations, Supplementary Table 9). The astacin family of metallopeptidases was the most expanded, with 184–387 copies in *Strongyloides*/*Parastrongyloides* compared with *Rhabditophanes* and with eight outgroup species, showing

that this expansion accompanies the evolution of parasitism (Fig. 1; Supplementary Table 10). Among the outgroup species the hookworm *Necator americanus*[38] has 82 astacin coding genes, and the free-living *C. elegans* 40[36].

SCP/TAPS proteins are often immunomodulatory molecules in parasitic nematodes[37] and have been investigated as potential vaccine candidates against *N. americanus*[39,40]. We found 89–205 SCP/TAPS coding genes in the *Strongyloides* spp. genomes, including nine subfamilies not present in *P. trichosuri*, *Rhabditophanes* or the eight outgroup species (Supplementary Tables 8 and 10). In *N. americanus* there are 137 SCP/TAPS coding genes[38], suggesting that this gene family has independently expanded twice: in nematode clades IV and V.

Additional gene expansions included receptor-type protein tyrosine phosphatases which have a putative role in signaling[41], and are expanded in *Strongyloides* and *Parastrongyloides* (52–75 genes) compared with *Rhabditophanes* (13), and the eight outgroup species (up to 39 genes). Acetylcholinesterase coding genes were expanded in *Strongyloides* and *Parastrongyloides* (30–126 genes) compared to *Rhabditophanes* (1) and 1–5 genes in our outgroup species. Many parasitic nematodes secrete acetylcholinesterases which are thought to facilitate their maintenance in hosts[42] and the expansion of this gene family in these parasitic species is consistent with this role. Some families show sub-clade specific expansion; for instance, *S. papillosus / S. venezuelensis* have a paralogous expansion of genes encoding Speckle-type POZ domains[43] (92–130 genes) compared with *S. ratti / S. stercoralis* (9–10 genes) (Fig. 1; Supplementary Table 8).

No function or annotation could be assigned to approximately one third (26–37%) of the genes present in the six species, but 50% of these could be assigned to novel gene families. The six largest of these families occurred only in *Strongyloides* and *Parastrongyloides*, comprising a total of 630 genes. We have named these *Strongyloides genome project families (sgpf) 1–6*. Members of s*gpf-1* and *-5* are predicted to have signal peptides and to be highly glycosylated (Supplementary Table 11).

Expanded gene familes are upregulated in parasitic stages

We identified genes and gene families that are likely to play a key role in the parasitic lifestyle of *S. ratti* and *S. stercoralis*, by comparing the transcriptomes of parasitic and free-living female stages. We generated *S. ratti* transcriptome data and used previously published *S. stercoralis* data[44]. A total of 909 *S. ratti* and 1,188 *S. stercoralis* genes were upregulated in parasitic females compared with free-living females (edgeR, fold change>2, FDR<0.01; Supplementary Tables 12, 13) of which 423 *S. ratti* and 457 *S. stercoralis* orthologous genes were upregulated in the parasitic female stage of both species (Supplementary Table 14).

The two most expanded *Strongyloides* gene families – SCP/TAPS[37] and astacin domain coding genes[45–48] – dominated the list of genes differentially expressed by the parasitic female. In *S. ratti* and *S. stercoralis*, respectively, 58 and 62% of putative astacin-like proteins and 57 and 71% SCP/TAPS genes were differentially expressed between parasitic vs. free-living females (Fig. 3; Supplementary Tables 10, 13). However, other paralogously expanded genes were not enriched among the upregulated genes suggesting they may not be

important for parasitism. Both *Strongyloides* and *Parastrongyloides* infect their hosts by skin penetration; the larvae then migrate through the host, and adult females in the host live in the mucosa of the small intestine[49,50] where they feed on the host. Astacins are metallopeptidases that have previously been associated with a role in tissue migration by nematode infective larvae[46,51]. Around half of the putative astacin-like proteins in *Strongyloides* spp. contain the canonical zinc binding motif (HEXXHXXGXXH) of astacin active sites and likely have a role in penetrating the host mucosa in which the parasitic females live. Teasing apart the role of different astacin gene family members in the migration and gut-dwelling phases of this life cycle could provide insights to allow new therapeutic interventions to be developed. For *S. ratti* and *S. stercoralis* respectively, 63 and 53% of the SCP/TAPS genes upregulated in the parasitic female encode a signal peptide suggesting that they may be secreted from the worm into the host. An immunomodulatory role for SCP/TAPS proteins has also been proposed based on the inhibitory effect that these proteins have on neutrophil and platelet activity in hookworm infections[37,52,53].

Other gene families commonly upregulated in the parasitic females of both species, compared with free-living females and iL3s, included those coding for transthyretin-like proteins, prolyl endopeptidases, acetylcholinesterases, trypsin-inhibitors, and aspartic peptidases (Fig. 3, Supplementary Table 15). The transthyretin-like genes had some of the highest fold changes of genes upregulated in the parasitic females (Supplementary Table 13). Transthyretin-like genes are a large, nematode-specific gene family[54], expressed in adult parasitic stages[55–57], and are distant relatives of vertebrate transthyretins that are involved in transporting thyroid hormones[58]. While some aspartic peptidases are essential for the digestion of host hemoglobin in blood-borne parasites[59,60], it has been proposed that others are involved in digesting other host macromolecules[61].

Hypothetical protein-coding genes accounted for 20–37% of the differentially expressed genes from pairwise comparisons of parasitic females, free-living females and iL3s, and included genes with the highest relative expression levels (Supplementary Table 13). These novel genes are likely to be important to these distinctive phases of the life cycle, including in parasitism. Three small novel gene families (*sgpf-7-9*) were predominantly upregulated in *S. ratti* parasitic females, two of which are predicted to be predominantly secretory or membrane-targeted (Supplementary Table 11). In contrast, the largest hypothetical protein-coding gene families, *sgpf-1–6*, accounted for only a small proportion (1% in both *S. ratti* and *S. stercoralis*) of all differentially expressed hypothetical protein-coding genes suggesting that they do not have roles involved in parasitism.

Using gene ontology annotations to summarize the putative functions of upregulated genes revealed distinct differences between the life cycle stages of both species (Fig. 3, Supplementary Table 16). The genes upregulated in iL3s appear to be associated with sensing the environment and with signal transduction, and were the most consistent between *S. ratti* and *S. stercoralis*. The products of free-living female expressed genes have core metabolic and growth-related roles (such as in cytoskeleton and chromatin). In parasitic stages, the dominant functional categories were proteases, consistent with the abundant astacins (Fig. 3, Supplementary Table 16).

The products of putative parasitism genes are secreted

In parallel we compared the somatic proteome of parasitic and free-living females of *S. ratti*. Of 1,266 proteins detected overall, 569 were comparatively upregulated in parasitic females and 409 in free-living females (Supplementary Tables 12, 17). We found a modest overlap between the transcriptome and somatic proteome; 6% of genes upregulated in the parasitic female transcriptome were also upregulated in the proteome, and 10% for free-living females (Supplementary Fig. 5; Supplementary Table 18). A poor concordance between transcript and peptide abundance has been reported in many systems[62–64] and likely reflects post-translational processes that decouple protein and mRNA abundance. In the present study, this may be compounded by the excretion / secretion of many gene products from parasitic stages, to interact with the host. Indeed, 43% of genes upregulated in the parasitic female transcriptome are predicted to encode signal peptides, compared with 26% for the free-living females. Furthermore, while several of the putative parasitism gene families were highly upregulated in the somatic proteome (aspartic peptidases, prolyl endopeptidases and acetylcholinesterases; Supplementary Table 17), we found only five astacin-like and no SCP/TAPS proteins (Supplementary Fig. 5). To address this we extended the analysis to the excretory/secretory (ES) proteome data of Soblik *et al*[65].

In the ES proteome we detected an additional 882 proteins, and found greater consistency with the parasitic female transcriptome: 13% of the parasitic female ES proteins overlapped with the upregulated transcriptome (Supplementary Table 18). We also found 25 astacin and 14 SCP/TAPS gene products in the ES proteome. Other gene families highly upregulated in the parasitic female transcriptome were also dominant in the parasitic ES proteome including prolyl endopeptidases, acetylcholinesterases, and transthyretin-like proteins (Supplementary Table 19). Protein products of novel gene families *sgpf*-1 and −5 were also identified in the ES products of both parasitic and free-living females (Supplementary Table 11). Other parasitic nematodes have been noted to have many protease coding genes, and different species appear to have expanded different protease families[38,66–68]. Together these, and our findings, suggest that expansion of protease coding genes, and secretion of extensive quantities of proteases is likely to be an essential feature of nematode parasitism. These proteases are, presumably, used to penetrate host tissue, acquire resources from the host and to protect the parasite from host-induced harm.

Parasitism-associated genes are in co-expressed clusters

We observed that genes upregulated in the parasitic females and iL3s were often physically clustered in the genome, more so than for genes upregulated in the free-living female (Supplementary Table 20). To test whether this clustering was significant we asked whether clusters of three or more adjacent genes, upregulated in the same life cycle stage, occurred more often than would be expected by chance. We found that 31%, 4% and 26% of upregulated genes were in such clusters in *S. ratti* parasitic females, free-living females and iL3s, respectively, while in *S. stercoralis* this was 34%, 2% and 34% (Supplementary Table 20). This clustering is more than would be expected by chance (Supplementary Fig. 6; Supplementary Table 20). The parasitic female clusters were larger (19 and 16 genes in the largest *S. ratti* and *S. stercoralis* clusters, respectively) compared with those of the iL3s (9 and 14 genes) and free-living female stages (3 genes) (Supplementary Table 20). Although

nematodes, including *S. ratti*[69], have operons these clusters are unlikely to be operons because (i) the average intergenic distance among clustered genes does not differ from the genome-wide average (Supplementary Fig. 6) and (ii) cluster members include genes on both strands.

Clusters of genes upregulated in the parasitic female were more likely to comprise genes from the same gene family. The majority (88–73 % for *S. ratti* and *S. stercoralis*, respectively) of these parasitic female clusters were of genes belonging to the same Compara gene family; this is greater than for iL3s (8–10%) (Supplementary Tables 20–22). Two gene families dominated parasitic female clusters: astacins (24 and 23% of parasitic female clusters for *S. ratti* and *S. stercoralis*) and SCP/TAPS (15 and 11%). Tandem expansions of astacin and SCP/TAPS genes could provide a plausible explanation for the preponderance of these gene families in the parasitic female expression clusters. However, even with the exclusion of the astacin and SCP/TAPS families, most remaining parasitic female clusters still comprised genes from the same gene family (85 and 65% for *S. ratti* and *S. stercoralis*, respectively); fewer clusters from the same gene family occurred for iL3s (7 and 9%) compared to parasitic females (Supplementary Table 21).

Phylogenetic analysis of astacins, including the eight outgroup species, showed that 139 *S. ratti* genes form one distinct clade (Fig. 4), presumably derived from a single ancestral astacin gene. Similarly, the *S. ratti* SCP/TAPS gene family has almost exclusively expanded from one ancestral gene (Fig. 4). These gene clusters likely arose by tandem duplication of genes, as has occurred for other large gene families, for example in *C. elegans*[18]. However, in contrast to *C. elegans*, physical adjacency of the duplicated genes has been maintained in *Strongyloides*, perhaps due to the expansions being recent and therefore not having yet been broken-up by recombination. Alternatively the adjacency may be functional, for example there being pressure to maintain a common regulatory environment. Clustering of gene families was relatively rare among *Rhabditophanes* and eight outgroup species (Supplementary Table 21), meaning that this clustering is specific to the *Strongyloides*/*Parastrongyloides* lineage and thus to the parasitic lifestyle in this clade.

The clusters of genes upregulated in the parasitic females were themselves chromosomally clustered forming 'parasitism regions' (Fig. 4). In *S. ratti* a third of genes upregulated in the parasitic female are concentrated in three regions of chromosome II, most notably a 3.6 Mb region at one end of chromosome II, comprising 171 genes that were upregulated in the parasitic female transcriptome (Supplementary Fig. 2). A similar pattern is evident in *S. stercoralis* where seven scaffolds and contigs with a high density of genes upregulated in the parasitic female also belong to chromosome II; 46% of the 171 *S. ratti* genes belong to just eight different gene families including those coding for aspartic peptidases, astacin-like, SCP/TAPS, transthyretin-like and trypsin inhibitor-like proteins. This is the first report of chromosomal clustering of genes likely to be important in nematode parasitism and hints at possible regulatory mechanisms for parasite development.

Discussion

Understanding the molecular and genetic differences between parasitic and free-living organisms is of fundamental biological interest, and essential to identify novel drug targets, and other methods to control parasitic nematodes and the diseases that they cause. We have undertaken a comparative genomics study of six taxa from an evolutionary clade that transitions from a free-living to parasitic lifestyle, which we combined with transcriptomic and proteomic analyses of parasitic and free-living female stages of *Strongyloides* spp. Together, this is a powerful way to discover the molecular adaptations to parasitism among these nematodes. We find that a preponderance of genes expanded in parasitic species are specifically used in the parasitic stages and are within genomic clusters, concentrated in regions of chromosome II. This is consistent with the idea that the within-host stages of parasitic nematodes deploy a specific biology that enables them to be successful parasites. The *Strongyloides* proteome and transcriptome have a limited overlap, as has been observed in other systems. For the *Strongyloides* clade we find that astacin and SCP/TAPS coding genes are prominent amongst parasitism-associated genes. Other parasitic nematodes appear to have expanded the number of protease coding genes in their genome, which also appear to be used predominantly during the within-host stages. In *Strongyloides* we have also found genomic clustering of these and other likely parasitism-associated genes, which is likely to have been initiated during the adaptation to parasitism, followed by subsequent repeated gene duplication, associated with adaptation to different hosts. This genomic arrangement may facilitate expression of a parasitic transcriptional program by these parasites. Operons have been demonstrated in *Strongyloides*, and it will be important to determine whether these parasitism associated genes are under operonic control.

Strongyloides is a particularly amendable laboratory system – both *S. ratti* and *S. venezuelensis* can be laboratory maintained in their natural rat host, as well as other rodents, and the parasite of humans *S. stercoralis* can also be maintained in the laboratory. In addition to providing a compelling model of the evolution of parasitism, transgenesis of *Strongyloides* and *Parastrongyloides* is possible[70–73] uniquely among parasitic nematodes, which will allow functional genomic studies, directed by our findings, to further explore the genetic basis of nematode parasitism.

Online Methods

Parasite material, sequencing and assembly

S. ratti, *S. stercoralis*, *S. venezuelensis* and *S. papillosus* larvae were obtained from fecal cultures of infected laboratory animals; for *Parastrongyloides trichosuri* and *Rhabditophanes* sp. KR3021 material was obtained from stages grown on agar plates. To produce the *S. ratti* reference genome, a combination of Sanger capillary, 454 and Illumina-derived sequence data was used, while data for the other species were generated using Illumina technology. The *S. ratti* genome was initially assembled using Newbler v.2.3[74] (for the capillary and 454 sequence data) and AbySS v.1.3.1[75] (for the Illumina data); Illumina paired-end reads were mapped to this with SMALT (Hannes Ponstingl, pers. comm.). The genomes of the other species, except *S. venezuelensis*, were assembled using a combination of SGA assembler[76] and Velvet[77], from 100 bp paired-end Illumina reads, produced from short (~500 bp)

fragment[78] and 3 kb mate-pair libraries[79]. Illumina reads were used in the IMAGE[80] and Gapfiller[81] software to fill gaps, and in iCORN[82] to correct base errors. Gap5[83] was used to manually extend and link scaffolds using Illumina read pairs. Genetic markers[22] were mapped to the *S. ratti* assembly to order and orient scaffolds, and in *S. papillosus* to assign scaffolds to chromosomes and regions of putative chromosomal diminution. The *S. venezuelensis* genome was assembled using the Platanus assembler[84] and improved as described above for other species. The resulting v2 *S. venezuelensis* assembly was further scaffolded using an optical map produced using an Argus optical mapping platform (Opgen). CEGMA v2[85] was used to assess the completeness of each assembly.

Assembled sequences were scanned for contamination from other species, using a series of BLASTX and BLASTP[86] searches against vertebrate and invertebrate sequence databases. Repeat sequences in the assemblies were characterized using RepeatModeler and TransposonPSI.

Mitochondrial genomes were assembled using MITObim assembler[87] with the *C. elegans* mitochondrial genes as seeds. The gene order of each assembly was confirmed by PCR. A mitochondrial protein-coding gene sequence phylogeny was constructed using RaxML v7.2.8[88].

Identifying regions that undergo chromatin diminution or belong to the X chromosome

To identify chromosomal regions that undergo chromatin diminution in *S. papillosus*, and scaffolds that belong to the X chromosome in *S. ratti*, *S. stercoralis*, and *P. trichosuri*, DNA of males and females from each species was sequenced and mapped to the appropriate reference genome using SMALT v0.7.4 (Hannes Ponstingl, pers. comm.. The read depth was calculated for each scaffold using the BedTools function genomecov[89], and all scaffolds were classified as diminished/X or non-diminished/autosomal based on differences in read coverage. Since males are hemizygous for the diminished region in *S. papillosus*[22], and for the X chromosome in the other species, a male: female read-depth ratio of 0.5:1 was expected in diminished or X scaffolds relative to autosomes, whereas in non-diminished/ autosomal region the ratio would be expected to be close to 1:1

Gene prediction and functional annotation

Genes were predicted using Augustus[90] – with a training set of approximately 200–400 manually curated genes per species, aligned transcript data and *S. ratti* protein sequences as hints – supplemented with non-overlapping predictions from MAKER[91]. If there was more than one alternative splice pattern for a gene prediction in the combined Augustus/MAKER gene set we only kept the transcript corresponding to the longest predicted protein. Astacin gene models and a subset of SCP/TAPS gene models from *S. ratti*, *S. venezuelensis* and *S. stercoralis* were manually curated prior to phylogenetic analyses.

A protein name was assigned to each predicted protein based on manually curated orthologs in UniProt[92] from selected species (human, zebrafish, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Schistosoma mansoni* orthologs) where possible. If a predicted protein was not assigned a protein name based on its orthologs, then a protein name was assigned based on InterPro[93] domains in the protein.

Gene Ontology (GO) terms were assigned by transferring GO terms from human, zebrafish, *C. elegans*, and *D. melanogaster* orthologs using an approach based on the Ensembl Compara approach for transferring GO terms to orthologs in vertebrate species[35], but modified for improved accuracy in transferring GO terms across phyla. Manually curated GO annotations were downloaded from the GO Consortium website[94], and for a particular predicted protein in the present study, the manually curated GO terms were obtained for all its human, zebrafish, *C. elegans*, and *D. melanogaster* orthologs. From this set the last common ancestor term (in the GO hierarchy) was found for each pair of GO terms from orthologs of two different species (*e.g.* a *C. elegans* ortholog and a zebrafish ortholog) and then transferred to our predicted protein. GO terms of the three possible types (molecular function, cellular component and biological process) were assigned to predicted proteins in this way. Additional GO terms were identified using InterproScan[95].

Gene orthology and species tree reconstruction

Eight outgroup species were used, encompassing four previously defined nematode clades[11] (clade I, *Trichinella spiralis, Trichuris muris*; clade III, *Ascaris suum, Brugia malayi*; clade IV, *Bursaphelenchus xylophilus, Meloidogyne hapla*; clade V, *Necator americanus, C. elegans*), together with the six species from the present study to construct a Compara database using the Ensembl Compara pipeline[35]. The database was used to identify orthologs and paralogs; gene duplications and gene losses; as well as gene families shared among the species, or sub-sets of the species, or specific to one species.

4,437 gene families were identified that contained just one gene from each species and that were present in at least ten species out of the six species and the eight outgroups. An alignment for the proteins in each family was built using MAFFT version v6.857[96], poorly-aligning regions were trimmed using GBlocks v0.91b, and the remaining columns were concatenated. For each alignment, the best-fitting amino acid substitution model was identified as that minimising the Akaike Information Criterion from the set of models available in RAxML v8.0.24[88], testing models with both pre-defined amino acid frequencies and observed frequencies in the data, and all with the CAT model of rate variation across sites. A maximum likelihood phylogenetic tree was constructed based on the concatenated alignment, with each protein alignment an independent partition of these data, applying the best-fitting substitution model identified above to each partition. This inference used RAxML v8.0.24 with ten random addition-sequence replicates and 100 bootstrap replicates, and otherwise default heuristic search settings.

Analysis of intron-exon structure and synteny analysis

Introns that were present in two or more species were identified from gene structures and full gene nucleotide alignments of 208 single-copy orthologs using ScipPio[97] and GenePainter[98]. The output from GenePainter was parsed into DOLLOP (PHYLIP package; Felsenstein, J.) to infer intron gain and loss on every node of the species tree using maximum parsimony.

Whole-assembly nucleotide alignments were produced between *S. ratti* and the other five species using nucmer[99]. Each scaffold from the other species was assigned a chromosome

based on its nucmer alignment to a *S. ratti* chromosome. To identify syntenic regions, conserved blocks of three consecutive orthologous genes or more in the same order and orientation were defined by DAGchainer[100], between the *S. ratti* reference and each of the other five species. To gain a high-level view of synteny, PROmer[101] was used to identify very highly conserved sequence matches, based on translated sequence, after which scaffolds from a particular species were ordered by matching to *S. ratti* chromosome and position in that chromosome, and the matches plotted using Circos[102].

Transcriptome and proteome analyses

For *S. ratti* and *S. stercoralis* the transcriptomes were compared from the parasitic female, free-living female and third stage infective larvae (iL3s); we note that parasitic and free-living adult females will have eggs *in utero*. For *S. ratti*, free-living females were picked individually from cultures of *S. ratti*-infected rat faeces, from where iL3s were also collected; parasitic females were collected by dissection of *S. ratti*-infected rats[103]. Two biological replicates were collected for parasitic and free-living females. These samples were divided approximately equally and used for both transcriptomic and proteomic analysis. A single biological sample was used for iL3 transcriptomic analysis. RNA was prepared from Trizol, and poly(A)RNA selected with Dynabeads, acoustically sheared and reverse transcribed to construct Illumina libraries that were sequenced. For *S. stercoralis* we used previously published data[44]. RNA-seq data were analyzed using R v.3.0.2 and the bioconductor package edgeR[104] to identify genes differentially expressed between all pairwise combinations of the three life-cycle stages.

For *S. ratti* the proteome was also compared between the parasitic and free-living females. Equivalent samples of the material collected for the transcriptome analyses were used. Protein was extracted by freeze / thawing, mechanical grinding and chemical extraction and digested with trypsin. The resulting peptide mixture was analyzed by liquid chromatography-mass spectrometry. Proteins were identified and quantified using Progenesis. For downstream analyses at least two unique peptides were required to identify proteins. Protein abundance (iBAQ) was calculated from Progenesis.

For both the transcriptome and proteome data, GO analysis was performed in R using TopGo v.2.16.0 and Fisher's exact test.

For the analysis of the ES proteome[65], converted raw spectral files were analysed by the Mascot search engine, where <1% FDR and a minimum of two significant peptides were required to identify proteins. Protein abundance was calculated from Mascot algorithm emPAI.

Astacins and SCP/TAPS

Genes encoding astacins and SCP/TAPS were identified using Interproscan. For these gene families we aligned amino acid sequences of all *S. ratti* and eight outgroup species' members using MAFFT[96]. The alignments were edited with TCS[105] using the weighted option and the distance matrix of the new alignment was calculated using ProtTest[106]. The phylogenetic tree was constructed by maximum likelihood using RAxML[88] with 100 bootstrap replicates.

Gene clusters

Clusters of genes were identified as three or more adjacent genes upregulated in the same stage of the life cycle. The members of a cluster were considered to share a common gene family where ≥50 % of genes belonged to the same Compara gene family. To investigate the number of clusters expected by chance for a particular life cycle stage, for *n* genes upregulated in a particular stage, we randomly selected *n* genes from the genome, and calculated the number of clusters seen for the *n* random genes; this was repeated 1000 times and the mean value calculated.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank from the WTSI: Coline Griffiths, David Willey, Richard Rance and DNA Pipelines; Jacqueline Keane and Daria Gordon for bioinformatics support; Matthew Dunn for the *S. venezuelensis* optical map; Anne Babbage for laboratory support; Magdalena Zarowiecki for gene finding and functional annotation advice. We thank for technical help Louise Hughes and Laura Weldon (University of Bristol); Holman Massey, Jr., Xinshe Li and Hongguang Shao (University of Pennsylvania); Dana. K. Howe and Riana I. Wernick (Oregon State University); Hubert Denise (European Bioinformatics Institute); Mitsuru Yabana (Tokyo Institute of Technology); Akina Hino and Ryusei Tanaka (University of Miyazaki) and Atsushi Toyoda (National Institute of Genetics) for sequencing. The *S. ratti* transcriptome and proteome work was funded by Wellcome Trust grant 094462/Z/10/Z awarded to M.V., J.W. and M.B. The *S. ratti, S. stercoralis, S. papillosus, P. trichosuri* and *Rhabditophanes* sp. genome sequencing and the *S. venezuelensis* optical mapping was funded by Wellcome Trust grant 098051. The *S. venezuelensis* work was supported by JSPS KAKENHI (Nos. 24310142, 21590466 and 24780044), KAKENHI for Innovative Areas “Genome Science” (No. 221S0002) and the Integrated Research Project for Human and Veterinary Medicine of the University of Miyazaki. I.J.T. was supported by Academia Sinica. Work was funded by grants AI050668 and AI105856 from the US National Institutes of Health (NIH) to J.B.L., and by Resource-related Research Grant RR02512 from N.I.H. to Mark Haskins, which provided research materials for the study. J.D.S. received support from NIH training grant AI060516. AK was supported by a pre-doctoral stipend from the Max Planck society. Work by AK, DH and AS was funded by the Max Planck Society.

References

1. Savioli L, Albonico M. Soil-transmitted helminthiasis. Nat. Rev. Microbiol. 2004; 2:618–619. [PubMed: 15303271]
2. Pullan RL, Brooker SJ. The global limits and population at risk of soil-transmitted helminth infections in 2010. Parasit. Vectors. 2012; 5:81. [PubMed: 22537799]
3. WHO | Soil-transmitted helminthiases. at http://www.who.int/gho/neglected_diseases/soil_transmitted_helminthiases/en/
4. WHO | Estimates for 2000–2012. at http://www.who.int/healthinfo/global_burden_disease/estimates/en/index2.html
5. Viney, ME.; Lok, JB. The biology of Strongyloides spp. Wormbook ed. The C. elegans Research Community. 2015. Wormbook doi/10.1895/wormbook.1.141.2, http://www.wormbook.org
6. Albonico M, Crompton DW, Savioli L. Control strategies for human intestinal nematode infections. Adv. Parasitol. 1999; 42:277–341. [PubMed: 10050275]
7. Crompton, DWT. Human helminthic populations. In Bailliere’s Clinical Tropical Medicine and Communicable Diseases. London: Academic Press; 1987.
8. Dorris M, Viney ME, Blaxter ML. Molecular phylogenetic analysis of the genus *Strongyloides* and related nematodes. Int. J. Parasitol. 2002; 32:1507–1517. [PubMed: 12392916]
9. Blaxter M, Koutsovoulos G, Jones M, Kumar S, Elsworth B. Phylogenomics of Nematoda. Syst. Assoc. Spec. Vol. Next Gener. Syst. In press.

10. Holovachov O, et al. A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. Nematology. 2009; 11:927–950.
11. Blaxter ML, et al. A molecular evolutionary framework for the phylum Nematoda. Nature. 1998; 392:71–75. [PubMed: 9510248]
12. Viney ME. A genetic analysis of reproduction in *Strongyloides ratti*. Parasitology. 1994; 109(4): 511–515. [PubMed: 7800419]
13. Viney ME, Matthews BE, Walliker D. Mating in the nematode parasite *Strongyloides ratti*: proof of genetic exchange. Proc. R. Soc. Lond. Ser. B. 1993; 254:213–219.
14. Tindall NR, Wilson PAG. An extended proof of migration routes of immature parasites inside hosts: pathways of *Nippostrongylus brasiliensis* and *Strongyloides ratti* in the rat are mutually exclusive. Parasitology. 1990; 100:281–288. [PubMed: 2345662]
15. Mackerras M. *Strongyloides* and *Parastrongyloides* (Nematoda: Rhabdiasoidea) in Australian Marsupials. Aust. J. Zool. 1959; 7:87.
16. Grant WN, et al. *Parastrongyloides trichosuri*, a nematode parasite of mammals that is uniquely suited to genetic analysis. Int. J. Parasitol. 2006; 36:453–466. [PubMed: 16500655]
17. Nemetschke L, Eberhardt AG, Viney ME, Streit A. A genetic map of the animal-parasitic nematode *Strongyloides ratti*. Mol. Biochem. Parasitol. 2010; 169:124–127. [PubMed: 19887089]
18. *C.elegans* sequencing consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science. 1998; 282:2012–2018. [PubMed: 9851916]
19. Hillier LW, et al. Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. PLoS Biol. 2007; 5:e167. [PubMed: 17608563]
20. Foth BJ, et al. Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction. Nat. Genet. 2014; 46:693–700. [PubMed: 24929830]
21. Kikuchi T, et al. Genomic insights into the origin of parasitism in the emerging plant pathogen *Bursaphelenchus xylophilus*. PLoS Pathog. 2011; 7:e1002219. [PubMed: 21909270]
22. Nemetschke L, Eberhardt AG, Hertzberg H, Streit A. Genetics, chromatin diminution, and sex chromosome evolution in the parasitic nematode genus *Strongyloides*. Curr. Biol. 2010; 20:1687–1696. [PubMed: 20832309]
23. Hino A, et al. Karyotype and reproduction mode of the rodent parasite *Strongyloides venezuelensis*. Parasitology. 2014; 141:1736–1745. [PubMed: 25089654]
24. Kulkarni A, Dyka A, Nemetschke L, Grant WN, Streit A. *Parastrongyloides trichosuri* suggests that XX/XO sex determination is ancestral in Strongyloididae (Nematoda). Parasitology. 2013; 140:1822–1830. [PubMed: 23953590]
25. Harvey SC, Viney ME. Sex determination in the parasitic nematode *Strongyloides ratti*. Genetics. 2001; 158:1527–1533. [PubMed: 11514444]
26. Hu M, Chilton NB, Gasser RB. The mitochondrial genome of *Strongyloides stercoralis* (Nematoda) - idiosyncratic gene order and evolutionary implications. Int. J. Parasitol. 2003; 33:1393–1408. [PubMed: 14527522]
27. Armstrong MR, Blok VC, Phillips MS. A multipartite mitochondrial genome in the potato cyst nematode *Globodera pallida*. Genetics. 2000; 154:181–192. [PubMed: 10628979]
28. Gibson T, et al. The mitochondrial subgenomes of the nematode *Globodera pallida* are mosaics: evidence of recombination in an animal mitochondrial genome. J. Mol. Evol. 2007; 64:463–471. [PubMed: 17479345]
29. Lunt DH, Hyman BC. Animal mitochondrial DNA recombination. Nature. 1997; 387:247. [PubMed: 9153388]
30. Humphreys-Pereira DA, Elling AA. Mitochondrial genome plasticity among species of the nematode genus *Meloidogyne* (Nematoda: Tylenchina). Gene. 2015; 560:173–183. [PubMed: 25655462]
31. Piganeau G. A broad survey of recombination in animal mitochondria. Mol. Biol. Evol. 2004; 21:2319–2325. [PubMed: 15342796]

32. Ballard JWO, Whitlock MC. The incomplete natural history of mitochondria. Mol. Ecol. 2004; 13:729–744. [PubMed: 15012752]
33. Sultana T, et al. Comparative analysis of complete mitochondrial genome sequences confirms independent origins of plant-parasitic nematodes. BMC Evol. Biol. 2013; 13:12. [PubMed: 23331769]
34. Sun L, Zhuo K, Lin B, Wang H, Liao J. The complete mitochondrial genome of *Meloidogyne graminicola* (Tylenchina): a unique gene arrangement and its phylogenetic implications. PLoS One. 2014; 9:e98558. [PubMed: 24892428]
35. Vilella AJ, et al. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 2009; 19:327–335. [PubMed: 19029536]
36. Park J-O, et al. Characterization of the astacin family of metalloproteases in *C. elegans*. BMC Dev. Biol. 2010; 10:14. [PubMed: 20109220]
37. Cantacessi C, et al. A portrait of the 'SCP/TAPS' proteins of eukaryotes - developing a framework for fundamental research and biotechnological outcomes. Biotechnol. Adv. 27:376–388. [PubMed: 19239923]
38. Tang YT, et al. Genome of the human hookworm *Necator americanus*. Nat. Genet. 2014; 46:261–269. [PubMed: 24441737]
39. Bethony JM, et al. Randomized, placebo-controlled, double-blind trial of the Na-ASP-2 hookworm vaccine in unexposed adults. Vaccine. 2008; 26:2408–2417. [PubMed: 18396361]
40. Goud GN, et al. Expression of the *Necator americanus* hookworm larval antigen Na-ASP-2 in *Pichia pastoris* and purification of the recombinant protein for use in human clinical trials. Vaccine. 2005; 23:4754–4764. [PubMed: 16054275]
41. Lemmon MA, Schlessinger J. Cell signaling by receptor tyrosine kinases. Cell. 2010; 141:1117–1134. [PubMed: 20602996]
42. Selkirk ME, Lazari O, Matthews JB. Functional genomics of nematode acetylcholinesterases. Parasitology. 2005; 131(Suppl):S3–S18. [PubMed: 16569291]
43. Kwon JE, et al. BTB Domain-containing Speckle-type POZ Protein (SPOP) Serves as an Adaptor of Daxx for Ubiquitination by Cul3-based Ubiquitin Ligase. J. Biol. Chem. 2006; 281:12664–12672. [PubMed: 16524876]
44. Stoltzfus JD, Minot S, Berriman M, Nolan TJ, Lok JB. RNAseq analysis of the parasitic nematode *Strongyloides stercoralis* reveals divergent regulation of canonical dauer pathways. PLoS Negl. Trop. Dis. 2012; 6:e1854. [PubMed: 23145190]
45. 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*. Mol. Biochem. Parasitol. 2010; 174:101–108. [PubMed: 20670659]
46. Williamson AL, et al. *Ancylostoma caninum* MTP-1, an astacin-like metalloprotease secreted by infective hookworm larvae, is involved in tissue migration. Infect. Immun. 2006; 74:961–967. [PubMed: 16428741]
47. Lun HM, Mak CH, Ko RC. Characterization and cloning of metallo-proteinase in the excretory/secretory products of the infective-stage larva of *Trichinella spiralis*. Parasitol. Res. 2003; 90:27–37. [PubMed: 12743801]
48. Semenova SA, Rudenskaya GN. The astacin family of metalloproteinases. Biochem. Suppl. Ser. B Biomed. Chem. 2009; 3:17–32.
49. Maruyama H, El-Malky M, Kumagai T, Ohta N. Secreted adhesion molecules of *Strongyloides venezuelensis* are produced by oesophageal glands and are components of the wall of tunnels constructed by adult worms in the host intestinal mucosa. Parasitology. 2003; 126:165–171. [PubMed: 12636354]
50. Maruyama H, Yabu Y, Yoshida A, Nawa Y, Ohta N. A role of mast cell glycosaminoglycans for the immunological expulsion of intestinal nematode, *Strongyloides venezuelensis*. J. Immunol. 2000; 164:3749–3754. [PubMed: 10725734]
51. Gomez Gallego S, et al. Identification of an astacin-like metallo-proteinase transcript from the infective larvae of *Strongyloides stercoralis*. Parasitol. Int. 2005; 54:123–133. [PubMed: 15866474]

52. Moyle M, et al. A hookworm glycoprotein that inhibits neutrophil function is a ligand of the integrin CD11b/CD18. J. Biol. Chem. 1994; 269:10008–10015. [PubMed: 7908286]
53. Del Valle A, Jones BF, Harrison LM, Chadderdon RC, Cappello M. Isolation and molecular cloning of a secreted hookworm platelet inhibitor from adult *Ancylostoma caninum*. Mol. Biochem. Parasitol. 2003; 129:167–177. [PubMed: 12850261]
54. Parkinson J, et al. A transcriptomic analysis of the phylum Nematoda. Nat. Genet. 2004; 36:1259–1267. [PubMed: 15543149]
55. Saverwyns H, et al. Analysis of the transthyretin-like (TTL) gene family in *Ostertagia ostertagi* - comparison with other strongylid nematodes and *Caenorhabditis elegans*. Int. J. Parasitol. 2008; 38:1545–1556. [PubMed: 18571174]
56. Jacob J, Vanholme B, Haegeman A, Gheysen G. Four transthyretin-like genes of the migratory plant-parasitic nematode *Radopholus similis*: members of an extensive nematode-specific family. Gene. 2007; 402:9–19. [PubMed: 17765408]
57. Chehayeb JF, Robertson AP, Martin RJ, Geary TG. Proteomic analysis of adult *Ascaris suum* fluid compartments and secretory products. PLoS Negl. Trop. Dis. 2014; 8:e2939. [PubMed: 24901219]
58. Richardson SJ, Hennebry SC, Smith BJ, Wright HM. Evolution of the thyroid hormone distributor protein transthyretin in microbes *C. elegans*, and vertebrates. Ann. N. Y. Acad. Sci. 2005; 1040:448–451. [PubMed: 15891085]
59. Williamson AL, et al. Cleavage of hemoglobin by hookworm cathepsin D aspartic proteases and its potential contribution to host specificity. FASEB J. 2002; 16:1458–1460. [PubMed: 12205047]
60. Longbottom D, et al. Molecular cloning and characterisation of a putative aspartate proteinase associated with a gut membrane protein complex from adult *Haemonchus contortus*. Mol. Biochem. Parasitol. 1997; 88:63–72. [PubMed: 9274868]
61. Mello LV, O'Meara H, Rigden DJ, Paterson S. Identification of novel aspartic proteases from *Strongyloides ratti* and characterisation of their evolutionary relationships, stage-specific expression and molecular structure. BMC Genomics. 2009; 10:611. [PubMed: 20015380]
62. Foss EJ, et al. Genetic basis of proteome variation in yeast. Nat. Genet. 2007; 39:1369–1375. [PubMed: 17952072]
63. Haider S, Pal R. Integrated analysis of transcriptomic and proteomic data. Curr. Genomics. 2013; 14:91–110. [PubMed: 24082820]
64. Ghazalpour A, et al. Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet. 2011; 7:e1001393. [PubMed: 21695224]
65. Soblik H, et al. Life cycle stage-resolved proteomic analysis of the excretome/secretome from *Strongyloides ratti* - identification of stage-specific proteases. Mol. Cell. Proteomics. 2011; 10:M111.010157. [PubMed: 21964353]
66. Laing R, et al. The genome and transcriptome of *Haemonchus contortus*, a key model parasite for drug and vaccine discovery. Genome Biol. 2013; 14:R88. [PubMed: 23985316]
67. Schwarz EM, et al. The genome and developmental transcriptome of the strongylid nematode *Haemonchus contortus*. Genome Biol. 2013; 14:R89. [PubMed: 23985341]
68. Schwarz EM, et al. The genome and transcriptome of the zoonotic hookworm *Ancylostoma ceylanicum* identify infection-specific gene families. Nat. Genet. 2015; 47:416–422. [PubMed: 25730766]
69. Guiliano DB, Blaxter ML. Operon conservation and the evolution of trans-splicing in the phylum Nematoda. PLoS Genet. 2006; 2:e198. [PubMed: 17121468]
70. Shao H, et al. Transposon-mediated chromosomal integration of transgenes in the parasitic nematode *Strongyloides ratti* and establishment of stable transgenic lines. PLoS Pathog. 2012; 8:e1002871. [PubMed: 22912584]
71. Li X, et al. Successful transgenesis of the parasitic nematode *Strongyloides stercoralis* requires endogenous non-coding control elements. Int. J. Parasitol. 2006; 36:671–679. [PubMed: 16500658]
72. Li X, et al. Transgenesis in the parasitic nematode *Strongyloides ratti*. Mol. Biochem. Parasitol. 2011; 179:114–119. [PubMed: 21723330]
73. Grant WN, et al. Heritable transgenesis of *Parastrongyloides trichosuri*: a nematode parasite of mammals. Int. J. Parasitol. 2006; 36:475–483. [PubMed: 16500659]

74. Margulies M, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437:376–380. [PubMed: 16056220]
75. Simpson JT, et al. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009; 19:1117–1123. [PubMed: 19251739]
76. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012; 22:549–556. [PubMed: 22156294]
77. Zerbino DR. Using the Velvet de novo assembler for short-read sequencing technologies. Curr. Protoc. Bioinformatics. 2010; 11:11–15.
78. Kozarewa I, et al. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat. Methods. 2009; 6:291–295. [PubMed: 19287394]
79. Park N, et al. An improved approach to mate-paired library preparation for Illumina sequencing. Methods Next Gener. Seq. 2013; 1
80. Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol. 2010; 11:R41. [PubMed: 20388197]
81. Nadalin F, Vezzi F, Policriti A. GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformatics. 2012; 13(1):S8. [PubMed: 23095524]
82. Otto TD, Sanders M, Berriman M, Newbold C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. Bioinformatics. 2010; 26:1704–1707. [PubMed: 20562415]
83. Bonfield JK, Whitwham A. Gap5-editing the billion fragment sequence assembly. Bioinformatics. 2010; 26:1699–1703. [PubMed: 20513662]
84. Kajitani R, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 2014; 24:1384–1395. [PubMed: 24755901]
85. Parra G, Bradnam K, Ning Z, Keane T, Korf I. Assessing the gene space in draft genomes. Nucleic Acids Res. 2009; 37:289–297. [PubMed: 19042974]
86. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]
87. Hahn C, Bachmann L, Chevreux B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - a baiting and iterative mapping approach. Nucleic Acids Res. 2013; 41:e129. [PubMed: 23661685]
88. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30:1312–1213. [PubMed: 24451623]
89. Quinlan AR. BEDTools: The Swiss-Army tool for genome feature analysis. Curr. Protoc. Bioinformatics. 2014; 47:11.12.1–11.12.34. [PubMed: 25199790]
90. Stanke M, et al. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006; 34:W435–W439. [PubMed: 16845043]
91. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011; 12:491. [PubMed: 22192575]
92. UniProt: a hub for protein information. Nucleic Acids Res. 2014; 43:D204–D212. [PubMed: 25348405]
93. Mitchell A, et al. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 2014; 43:D213–D221. [PubMed: 25428371]
94. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2014; 43:D1049–D156. [PubMed: 25428369]
95. Jones P, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014; 30:1236–1240. [PubMed: 24451626]
96. Katoh K, Standley DM. MAFFT: iterative refinement and additional methods. Methods Mol. Biol. 2014; 1079:131–146. [PubMed: 24170399]
97. Keller O, Odronitz F, Stanke M, Kollmar M, Waack S. Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. BMC Bioinformatics. 2008; 9:278. [PubMed: 18554390]

98. Hammesfahr B, Odronitz F, Mühlhausen S, Waack S, Kollmar M. GenePainter: a fast tool for aligning gene structures of eukaryotic protein families, visualizing the alignments and mapping gene structures onto protein structures. BMC Bioinformatics. 2013; 14:77. [PubMed: 23496949]

99. Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 2002; 30:2478–2483. [PubMed: 12034836]

100. Haas BJ, Delcher AL, Wortman JR, Salzberg SL. DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics. 2004; 20:3643–3646. [PubMed: 15247098]

101. Kurtz S, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5:R12. [PubMed: 14759262]

102. Krzywinski M, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19:1639–1645. [PubMed: 19541911]

103. Thompson FJ, Barker GLA, Hughes L, Viney ME. Genes important in the parasitic life of the nematode *Strongyloides ratti*. Mol. Biochem. Parasitol. 2008; 158:112–119. [PubMed: 18234359]

104. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–140. [PubMed: 19910308]

105. Chang J-M, Di Tommaso P, Notredame C. TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. Mol. Biol. Evol. 2014; 31:1625–1637. [PubMed: 24694831]

106. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 2005; 21:2104–2105. [PubMed: 15647292]

107. Bolla RI, Roberts LS. Gametogenesis and chromosomal complement in *Strongyloides ratti* (Nematoda: Rhabdiasoidea). J. Parasitol. 1968; 54:849–855. [PubMed: 4919050]

108. Hammond MP, Robinson RD. Chromosome complement, gametogenesis, and development of *Strongyloides stercoralis*. J. Parasitol. 1994; 80:689–695. [PubMed: 7931903]

109. Albertson DG, Nwaorgu OC, Sulston JE. Chromatin diminution and a chromosomal mechanism of sexual differentiation in *Strongyloides papillosus*. Chromosoma. 1979; 75:75–87. [PubMed: 533664]

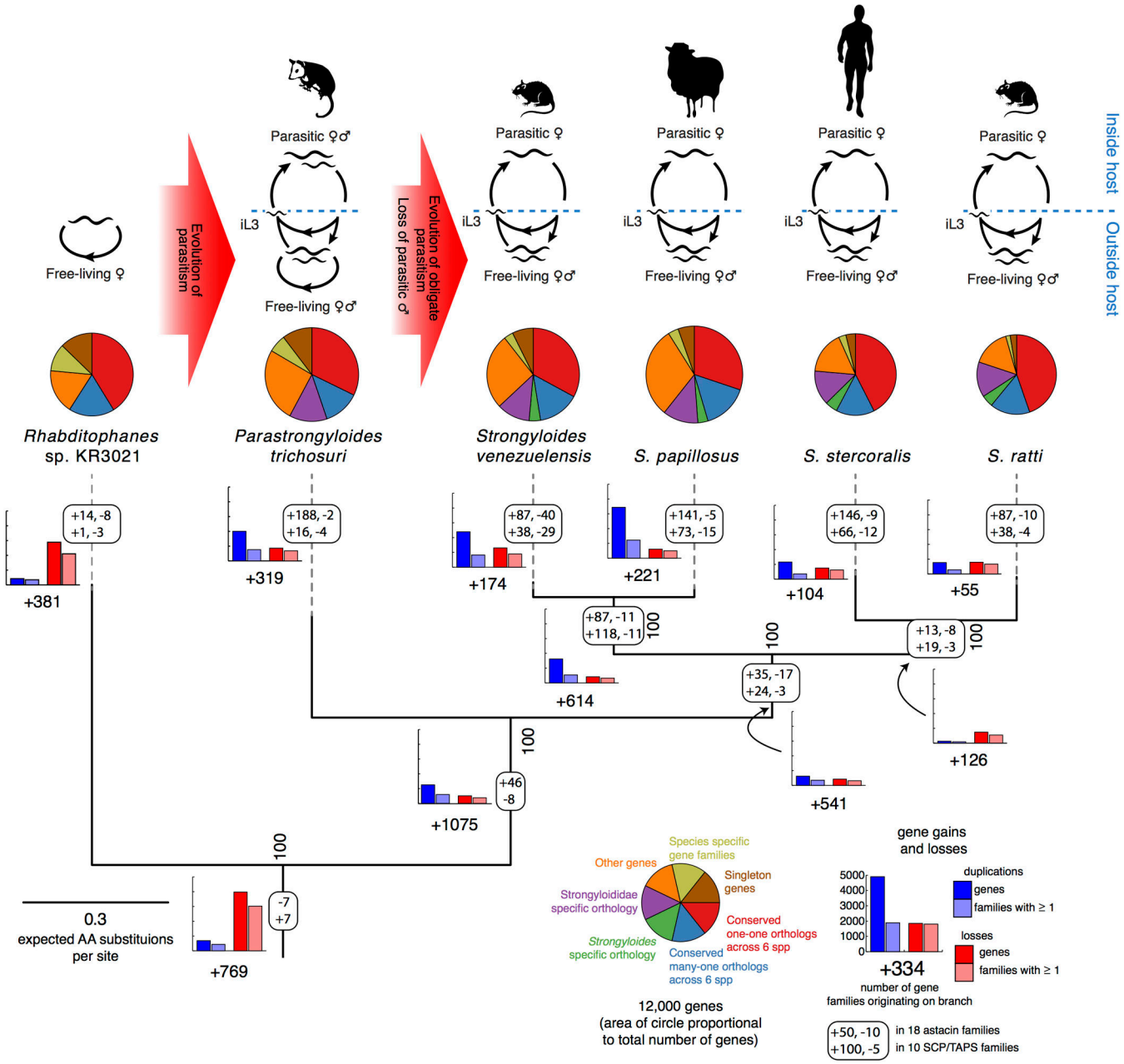

Fig. 1. Evolution and comparative genomics of *Strongyloides* and relatives

The life cycles of six clade IV nematodes showing the transition from a free-living lifestyle (in *Rhabditophanes*), through facultative parasitism (*P. trichosuri*), to obligate parasitism (*Strongyloides* spp.), and the phylogeny of these species (maximum-likelihood phylogeny based on a concatenated alignment of 841,529 amino acid sites from 4,437 conserved single-copy orthologous genes). Values on nodes (all 100) are the number of bootstrap replicate trees showing the split induced by the node, out of 100 bootstrap replicates. The phylogeny is annotated with the numbers of gene families appearing along each branch of the phylogeny (+values on each branch) and histograms show the number of duplications (blue) and losses (red) for individual genes (dark blue or red) and foor families (light blue or red);

the number of gene origins and gene losses in 18 astacin families (upper numbers in boxes) and ten SCP/TAPS families (lower numbers in boxes) as estimated by the Ensembl Compara pipeline is also shown. The pie charts summarize the evolutionary history of the genome of each species, defining genes shared among all six species, the five parasitic species (Strongyloididae, which includes all except *Rhabditophanes*), the four *Strongyloides* species, and species-specific genes. The host species of the parasites are shown: for *P. trichosuri* the brushtail possum, for *S. ratti* and *S. venezuelensis* the rat, for *S. stercoralis* humans, and for *S. papillosus* sheep.

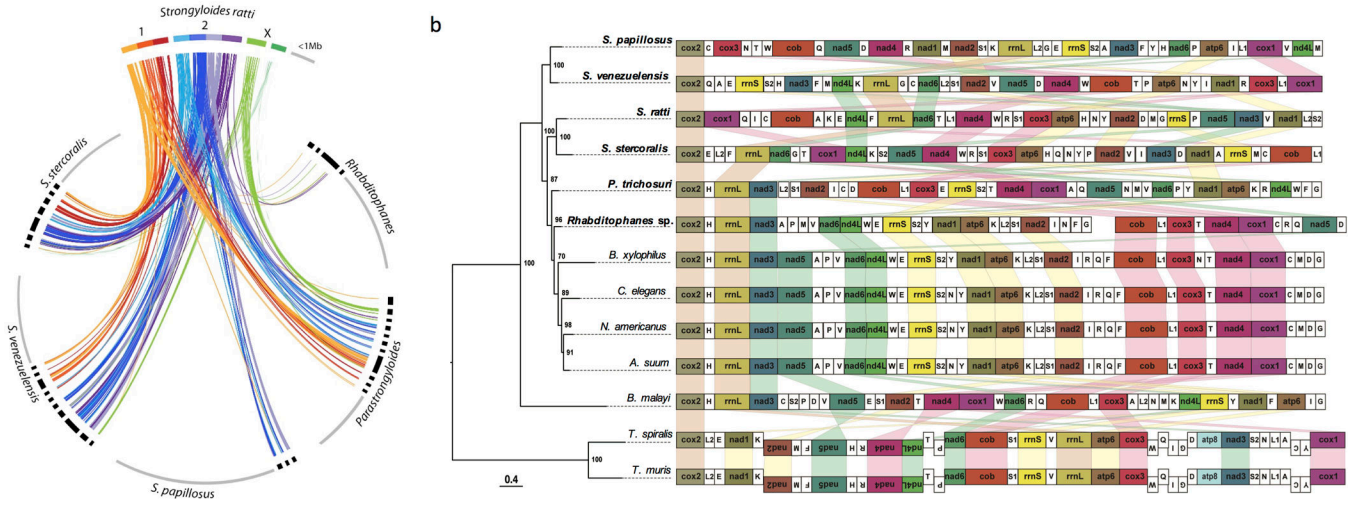

Fig. 2. Nuclear genomic synteny and mitochondrial genomes of four *Strongyloides* spp., *P. trichosuri* and *Rhabditophanes* sp

(a) The *S. ratti* genome, our best assembled genome, is used as the reference sequence; synteny is based on sequence matches. Graduation of color across the *S. ratti* chromosomes represents position along the chromosome for chromosome I (yellow-red), chromosome II (blue-purple) and chromosome X (green). Black boxes represent scaffolds >1Mb; scaffolds <1Mb are grouped together and shown in grey. (b) The mitochondrial gene order and phylogeny for our six species and seven outgroup species that encompass four nematode clades. Our eighth outgroup species, *Meloidogyne hapla*, was excluded due to insufficient mitochondrial genome data. Inverted sequences are shown by gene boxes with inverted text. The maximum likelihood tree (left) was constructed using 12 mitochondrial proteins. Amino acid sequences were aligned before concatenation (Supplementary Note).

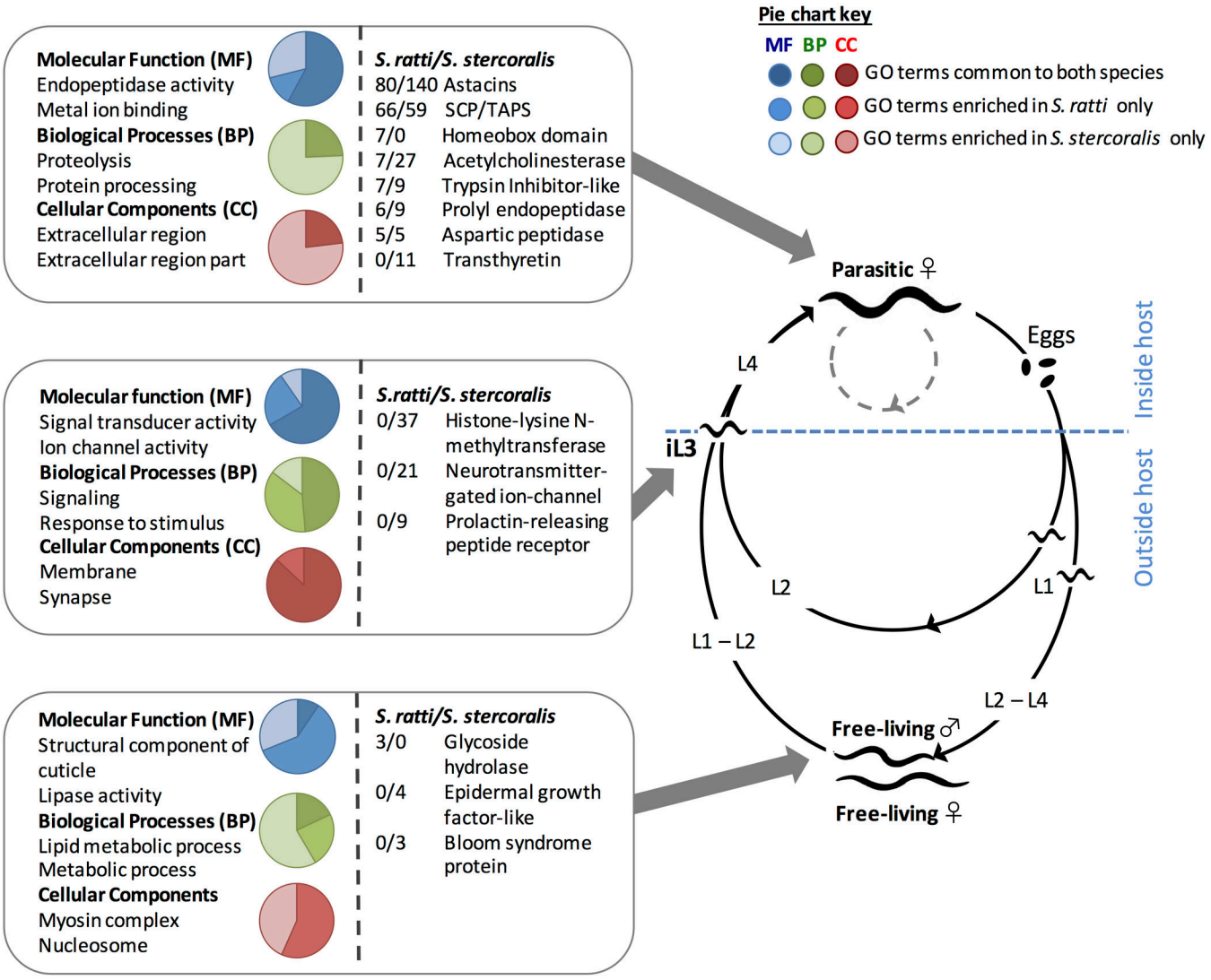

Fig. 3. The parasitic female, free-living female and infective third-stage larvae transcriptomes of *Strongyloides* spp

The progeny of the parasitic female pass out of the host (as larvae for *S. stercoralis*, or eggs and larvae for *S. ratti*) where infective third stage larvae (iL3s) can develop directly, or free-living males and females develop, whose progeny develop into iL3s; iL3s then infect hosts. The human parasite, *S. stercoralis*, can undergo internal auto-infection (grey dashed line) where iL3s develop and internally reinfect the same host. The transcriptome of the parasitic females, free-living females and iL3s were compared for *S. ratti* and *S. stercoralis*. Representative GO terms that were significantly enriched (left-hand side area of box) and Ensembl Compara gene families significantly upregulated (right-hand side of box) for each of these three stages of the lifecycle is summarized. The pie charts show the proportion of the GO terms common to *S. ratti* and *S. stercoralis*, or unique to either. Numbers in the right-hand side of boxes represent the number of genes upregulated in each gene family for *S. ratti* and *S. stercoralis*.

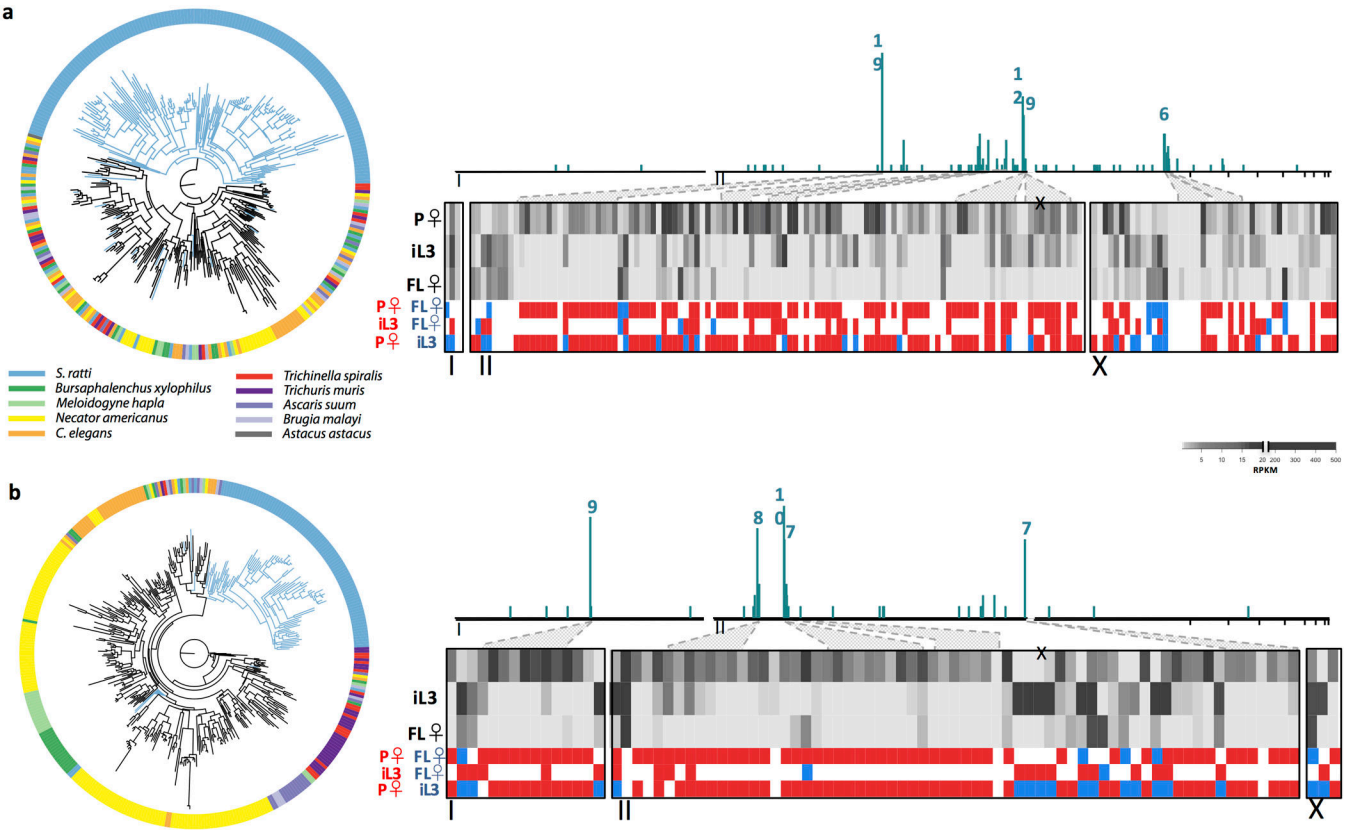

Fig. 4. *Strongyloides*-specific expansions and chromosomal clustering of gene families

(a) Astacin-like and (b) SCP/TAPS are the two major *Strongyloides ratti* gene families upregulated in the transcriptome of parasitic females. Left shows the phylogeny of each of these for *S. ratti* and our eight outgroup species and the crayfish *Astacus astacus S. ratti* genes are in light blue. Right shows the distribution of these genes in the genome, plotted as clusters of physically adjacent genes in the genome. Numbers above the peaks are the number of genes in a cluster of physically neighboring genes; ticks below the axis denote scaffold boundaries for chromosome X. The transcriptomic expression of these genes (in RPKM, reads per kilobase per million mapped reads) for parasitic females, free-living females and iL3s are shown on a grey scale, and the results of pairwise edgeR analysis of the gene expression among these lifecycle stages is shown in red or blue where a gene is upregulated. The color representing upregulation (red or blue) in a given stage of the life cycle relates to the color of the name of that stage for each pairwise comparison (fold change > 2, FDR < 0.01); no differential expression is shown as a white block.

Table 1

Properties of genome assemblies

Genome statistics based on scaffolds, excluding scaffolds less than 1000 bp. N50 is the size above which 50% of the assembled bases are distributed; N50 (number) is the number of scaffolds in which 50% of assembled bases exist.

	S. ratti	S. stercoralis	S. papillosus	S. venezuelensis	P. trichosuri	Rhabditophanes	C. elegans
Clade	IV	IV	IV	IV	IV	IV	V
Number of chromosomes	3^{107}	3^{108}	2^{109}	2^{22}	3^{23}	5^{a}	6^{18}
Assembly version	V5.0.4	V2.0.4	V2.1.4	V2.0.4	V2.0.4	V2.0.4	WS244
Assembly size (Mb)	43.1	42.6	60.2	52.1	42.2	47.2	100.2
Number of scaffolds	115^{b}	675	4,353	520	1,391	380	6
N50 of scaffolds (kb)	11,700	431	86	715	837	537	17,500
N50 (number)	2	16	129	16	12	22	3
Maximum scaffold length (Mb)	16.8	5.0	1.7	5.9	6.2	7.3	20.9
G+C content (%)	21	22	26	25	31	32	36
Number of genes	12,451	13,098	18,457	16,904	15,010	13,496	23,629
Number of exons	33,796	34,366	40,821	40,619	35,049	37,987	145,275
Exons, combined length (Mb)	17.5	17.9	22.4	20.3	20.8	17.8	30.1
Median exon length (bp)	263	265	304	261	348	276	146
Number of introns	21,345	21,268	22,364	23,715	20,039	24,491	169,506

[a] See Supplementary Figure 7

[b] 12 scaffolds, covering 93% of the genome, are assigned to chromosomes; 103 scaffolds are not assigned to a chromosome.