Comparative genomics of the major parasitic worms

International Helminth Genomes Consortium*

Parasitic nematodes (roundworms) and platyhelminths (flatworms) cause debilitating chronic infections of humans and animals, decimate crop production and are a major impediment to socioeconomic development. Here we report a broad comparative study of 81 genomes of parasitic and non-parasitic worms. We have identified gene family births and hundreds of expanded gene families at key nodes in the phylogeny that are relevant to parasitism. Examples include gene families that modulate host immune responses, enable parasite migration though host tissues or allow the parasite to feed. We reveal extensive lineage-specific differences in core metabolism and protein families historically targeted for drug development. From an in silico screen, we have identified and prioritized new potential drug targets and compounds for testing. This comparative genomics resource provides a much-needed boost for the research community to understand and combat parasitic worms.

Over a quarter of humans are infected with parasitic nematodes (roundworms) or platyhelminths (flatworms). Although rarely lethal, infections are typically chronic, leading to pain, malnutrition, physical disabilities, delayed development, deformity, social stigma or a burden on family members caring for the afflicted. These diseases encompass many of the most neglected tropical diseases and attract little research investment. Parasitic nematodes and platyhelminths impede economic development through human disability, and billions of dollars of lost production in the livestock and crop industries.

Few drugs are available to treat worm infections. Repeated mass administration of monotherapies is increasing the risk of resistance to human anthelmintics and has driven widespread resistance in farm animals. There are no vaccines for humans, and few for animals. The commonly used nematicides of plant parasites are environmentally toxic, and need replacement.

The phylum Nematoda is part of the superphylum Ecdysozoa and has five major clades (I to V), four of which contain human-infective parasites and are analyzed here (Fig. 1). The phylum Platyhelminthes is part of the superphylum Lophotrochozoa and the majority of parasite species are cestodes (tapeworms) and trematodes (flukes). Comparing the genomes of parasites from these two phyla may reveal common strategies employed to subvert host defenses and drive disease processes.

We have combined 36 published genomes with new assemblies for 31 nematode and 14 platyhelminth species into a large genome comparison of parasitic and non-parasitic worms. We have used these data to identify gene families and processes associated with the major parasitic groups. To accelerate the search for new interventions, we have mined the dataset of more than 1.4 million genes to predict new drug targets and drugs.

Results

Genomic diversity in parasitic nematodes and platyhelminths. We have produced draft genomes for 45 nematode and platyhelminth species and predicted 0.8 million protein-coding genes, with 9,132–17,274 genes per species (5–95% percentile range; see Methods, Supplementary Tables 1–3, Supplementary Fig. 1 and Supplementary Notes 1.1 and 1.2). We combined these new data with 36 published worm genomes—comprising 31 parasitic and five free-living species—and 10 outgroups from other animal phyla, into a comparative genomics resource of 91 species (Fig. 1 and Supplementary Tables 2 and 4). There was relatively little variation in gene set completeness (coefficient of variation, c.v. = 0.15) among the nematodes and platyhelminths, despite variation in assembly contiguity (c.v. = 8.5; Fig. 1b and Supplementary Table 2). Nevertheless, findings made using a subset of high-quality assemblies that were designated ‘tier 1’ (Methods and Supplementary Table 4) were corroborated against all species.

Gene family births and expansions. We inferred gene families from the predicted proteomes of the 91 species using Ensembl Compara. Of the 1.6 million proteins, 1.4 million were placed into 108,351 families (Supplementary Note 2.1 and Supplementary Data), for which phylogenetic trees were built and orthology and paralogy inferred (Methods, Supplementary Fig. 2 and Supplementary Table 7). Species trees inferred from 202 single-copy gene families that were present in at least 25% of species (Fig. 1), or from presence/absence of gene families, largely agreed with the expected species and clade relationships, except for a couple of known contentious issues (Supplementary Fig. 3, Supplementary Note 2.2 and Methods).

The species in our dataset contained significant novelty in gene content. For example, ~28,000 parasitic nematode gene families contained members from two or more parasitic species but were absent from Caenorhabditis elegans and 47% of gene families lacked any functional annotation (Supplementary Note 2.1 and Methods).

*A list of members and affiliations appears at the end of the paper.
Fig. 1 | Genome-wide phylogeny of 56 nematode, 25 platyhelminth species and 10 outgroup species. a. Maximum-likelihood phylogeny based on a partitioned analysis of a concatenated data matrix of 21,649 amino acid sites from 202 single-copy orthologous proteins present in at least 23 of the species. Values on marked nodes are bootstrap support values; all unmarked nodes were supported by 100 bootstrap replicates; nodes with solid marks were constrained in the analysis. Bar plots show genome sizes and total lengths of different genome features, and normalized gene count (Supplementary Note 1.2) for proteins with inferred functions based on sequence similarity (having an assigned protein name; Methods), or those without (named ‘hypothetical protein’). Species for which we have sequenced genomes are marked with asterisks; 33 ‘tier 1’ genomes are in black. LTR, long terminal repeat; LINE, long interspersed nuclear element. b. Assembly statistics. Blue rows indicate the 33 ‘tier 1’ genomes. Asterisks indicate the species for which we have sequenced genomes.
The latter families tended to be smaller than those with annotations (Supplementary Fig. 4) and, in many cases, correspond to families that are so highly diverged that ancestry cannot be traced, reflecting the huge breadth of unexplored parasite biology.

Gene families specific to particular parasite clades are likely to reflect important aspects of parasite biology and possible targets for new antiparasitic interventions. At key nodes in the phylogeny that are relevant to parasitism, we identified 5,881 families with apparent clade-specificity (sympomorphies; Supplementary Note 2.3, Methods and Supplementary Table 8), although our ability to discriminate truly parasite-specific clades was limited by the low number of free-living species. The apparent sympomorphies were either gene family births, or subfamilies that were so diverged from their homologues that they appeared as separate families. Functional annotation of these families was diverse (Fig. 2), but they were frequently associated with sensory perception (such as G-protein coupled receptors; GPCRs), parasite surfaces (platyhelminth tegument or nematode cuticle maintenance proteins) and protein degradation (proteases and protease inhibitors).

Among nematodes, clade IVa (which includes Strongylidae spp.; Fig. 1) showed the highest number of clade-specific families, including a novel ferrochelatase-like family. Most nematodes lack functional ferrochelatas for the last step of haem biosynthesis, but harbor ferrochelatase-like genes of unknown function, to which the sympomorphic clade IVa family was similar (Supplementary Fig. 5 and Methods). Exceptions are animal parasites in nematode clades III (for example ascarids and filaria) and IV that acquired a functional ferrochelatase via horizontal gene transfer. Within the parasite platyhelminths, a clade-specific inositol-pentakisphosphate 2-kinase (IP2K) was identified. In some species of Echinococcus tapeworms, IP2K produces inositol hexakisphosphate nanodeposits in the extracellular wall (the laminated layer) that protects larval metacestodes. The deposits increase the surface area for adsorption of host proteins and may promote interactions with the host.

Paralogous expansions of gene families, particularly those that are large or repeatedly involve related processes, can be evidence of adaptive evolution. We searched among our 10,986 highest-confidence gene families (those containing ≥10 genes from tier 1 species) for those that had expanded in parasite clades. A combination of scoring metrics (Methods) reduced the list to 995 differentially distributed families with a bias in copy number in at least one parasite clade. Twenty-five expansions have previously been observed, including 21 with possible roles in parasitism (Supplementary Fig. 6). A further 43 were placed into major functional classes that historically have been favored as drug targets (kinases, GPCRs, ion channels and proteases; Supplementary Table 9a).

By manually inspecting the distribution of the remaining 927 families across the full species tree, we identified 176 families with striking expansions (Supplementary Table 9a and Supplementary Note 2.4). Thirty two had no functional annotation; for example, family 393312 was highly expanded in clade Va nematodes (Supplementary Fig. 7 and Supplementary Table 9a). Even when families could be functionally annotated to some extent (for example, based on a protein domain), discerning their precise biological role was a challenge. For example, a sulfotransferase family that was expanded in flukes compared with tapeworms includes the Schistosoma mansoni locus that is implicated in resistance to the drug oxamniquine but the endogenous substrate for this enzyme is unknown (Supplementary Fig. 7j).

Among the newly identified expansions, we focused on those with richer functional information, especially where they were related to similar biological processes. For instance, we identified several expansions of gene families involved in innate immunity of the parasites, as well as their development. These included families implicated in protection against bacterial or fungal infections in nematode clade IVa (bus-4 GT31 galactosyltransferase, irg-3) and clades Va/Vc (lysozyme and the dual oxidase bli-3) (Supplementary Fig. 8a–d). In nematode clade IIIb, a family was expanded that contains orthologs of the Parascaris coiled-coiled protein PUMA, involved in kinetochore biology (Fig. 2b). This expansion possibly relates to the evolution of chromatin diminution in this clade, which results in an increased number of chromosomes requiring correct segregation during metaphase. In nematode clade IVa and in Bursaphelenchus, an expansion of a steroid kinase family (Supplementary Fig. 8e) is suggestive of novelty in steroid-regulated processes in this group, such as the switch between free-living or parasitic stages in Strongylidae.

Infections with parasitic worms are typified by their chronicity and a plausible involvement in host–parasite interactions is a recurring theme for many of the families. Taenia tapeworms and clade V strongylid nematodes (that is Va, Vb and Vc; Fig. 1) contained two expanded families with apryrase domains that may have a role in hydrolyzing ATP (a host danger signal) from damaged host tissues (Fig. 2b and Supplementary Fig. 9a). Moreover, many of the strongylid members also contained amine oxidoreductase domains, possibly to reduce production of pro-inflammatory amines, such as histamine, from host tissues. In platyhelminths, we observed expansions of tetraspanin families that are likely components of the host/pathogen interface. Described examples show tetraspanins being part of extracellular vesicles released by helminths within hosts; or binding the Fc domain of host antibodies; or being highly immunogenic (Supplementary Fig. 9b,c). In strongylids, especially clade Vc, an expansion of the fatty acid and retinol-binding (FAR) family, implicated in host–parasite interaction of plant- and animal-parasitic nematodes (Supplementary Fig. 9d), suggests a role in immune modulation. Repertoires of glycosyltransferases have expanded in nematode clades Vc and IV, and tapeworms (Supplementary Fig. 10a–c), and may be used to evade or divert host immunity by modifying parasite surface molecules directly exposed to the immune system; alternatively, surface glycoproteins may interact with lectin receptors on innate immune cells in an inhibitory manner. An expanded chondroitin hydrolase family in nematode clade Vc may possibly be used either for larval migration through host connective tissue or to digest host intestinal walls (Supplementary Fig. 9e). Similarly, an expanded GH5 glycosyl hydrolase family contained schistosome members with egg-enriched expression that may be used for traversing host tissues such as bladder or intestinal walls (Supplementary Fig. 9f). In nematode clade I, we found an expansion of a family with the PAN/ Apple domain, which is implicated in attachment of some protozoan parasites to host cells, and possibly modulates host lectin-based immune activation (Supplementary Fig. 9g).

The SCP/TAPS (sperm-coating protein/Tpx/antigen 5/pathogenesis-related protein 1) genes have been associated with parasitism through their abundance, secretion and evidence of their role in immunomodulation but are poorly understood. This diverse superfamly appeared as eight expanded Compara families. A more comprehensive phylogenetic analysis of the full repertoire of 3,167 SCP/TAPS sequences (Supplementary Note 2.5, Supplementary Table 10 and Methods) revealed intra- and interspecific expansions and diversification over different evolutionary timescales (Fig. 3 and Supplementary Figs. 11a,b and 12). In particular, the SCP/TAPS superfamly has expanded independently in nematode clade V (18–381 copies in each species) and in clade IVa parasites (39–166 copies) (Fig. 3 and Supplementary Fig. 11c). Dracunculus medinensis (Guinea worm) was unusual in being the only member of clade III to display an expansion (66 copies), which may reflect modulation of the host immune response during the tissue migration phase of its large adult females.

**Proteins historically targeted for drug development.** Proteases, GPCRs, ion channels and kinases dominate the list of targets.
Fig. 2 | Functional annotation of synapomorphic and expanded gene families. a, Rectangular matrices indicate counts of synapomorphic families grouped by 18 functional categories, detailed in the top left corner. Representative functional annotation of a family was inferred if more than 90% of the species present contained at least one gene with a particular domain. The node in the tree to which a panel refers is indicated in each matrix. ‘Other’ indicates families with functional annotation that could not be grouped into one of the 18 categories. ‘None’ indicates families that had no representative functional annotation. b, Expansions of apyrase and PUMA gene families. Families were defined using Compara. For color key and species labels, see Fig. 1. The plot for a family shows the gene count in each species, superimposed on the species tree. A scale bar beside the plot for a family shows the minimum, median and maximum gene count across the species, for that family.
such as larval development and host invasion. These expansions are particularly expanded in species that feed on blood (two expansions in clades Va and Vc; Supplementary Table 11). Because many of these species invade through skin (IVa, Vc; Supplementary Table 12) and migrate through the digestive system and lung (IVa, Vc, Vb; Supplementary Table 13), these expansions are consistent with evidence that astacins are involved in skin penetration and migration through connective tissue. The cathepsin B C1-cysteine proteases, which are strongly expressed in the parasitic stages and to a lesser extent in clade IIIb (Fig. 4), are involved in skin penetration and migration.

Independent expansion and functional divergence has differentiated the nematode and platyhelminth pentameric ligand gated ion channels (Supplementary Fig. 16, Supplementary Table 16 and Supplementary Note 3.6). All parasitic nematodes possessed chemoreceptors, with the most in clade IVa, including several large families synapomorphic to this clade (Supplementary Fig. 15), perhaps related to their unusual life cycles that alternate between free-living and parasitic forms. Independent expansion and functional divergence has differentiated the nematode and platyhelminth pentameric ligand gated ion channels (Supplementary Fig. 16, Supplementary Table 16 and Supplementary Note 3.6). All parasitic nematodes possessed chemoreceptors, with the most in clade IVa, including several large families synapomorphic to this clade (Supplementary Fig. 15), perhaps related to their unusual life cycles that alternate between free-living and parasitic forms.

Independent expansion and functional divergence has differentiated the nematode and platyhelminth pentameric ligand gated ion channels (Supplementary Fig. 16, Supplementary Table 16 and Supplementary Note 3.6). All parasitic nematodes possessed chemoreceptors, with the most in clade IVa, including several large families synapomorphic to this clade (Supplementary Fig. 15), perhaps related to their unusual life cycles that alternate between free-living and parasitic forms. Independent expansion and functional divergence has differentiated the nematode and platyhelminth pentameric ligand gated ion channels (Supplementary Fig. 16, Supplementary Table 16 and Supplementary Note 3.6). All parasitic nematodes possessed chemoreceptors, with the most in clade IVa, including several large families synapomorphic to this clade (Supplementary Fig. 15), perhaps related to their unusual life cycles that alternate between free-living and parasitic forms.
A range of annotated enzymes per species than the platyhelminths (Supplementary Fig. 20a), in part reflecting the paucity of biochemical studies in platyhelminths. Because variation in assembly quality or divergence from model organisms could bias enzyme predictions, we identified losses of pathways and differences in pathway coverage across different clades (Supplementary Note 4, Methods, Other Nematoda Phylogeny Species).  

**Fig. 4** Abundances of superfamilies historically targeted for drug development. Relative abundance profiles for 84 protease and 31 protease inhibitor families represented in at least 3 of the 81 nematode and platyhelminth species. Thirty-three protease families and 6 protease inhibitor families present in fewer than 3 species were omitted from the visualization. For each species, the gene count in a class was normalized by dividing by the total gene count for that species. Families mentioned in the Results or Supplementary Note text are labeled; complete annotations of all protease families are in Supplementary Table 11.
Fig. 5 and Supplementary Fig. 21). Pathways related to almost all metabolic superpathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG)\(^\text{86}\) showed significantly lower coverage for platyhelminths (versus nematodes) and filaria (versus other nematodes) (Supplementary Fig. 20b).

In contrast to most animals, nematodes possess the glyoxylate cycle that enables conversion of lipids to carbohydrates, to be used for biosyntheses (for example, during early development) and to avert starvation\(^\text{37}\). The glyoxylate cycle appears to have been lost independently in the filaria and *Trichinella* species (Fig. 5a; M00012), both of which are tissue-dwelling obligate parasites. The filaria and *Trichinella* have also independently lost alanine-glyoxylate transaminase that converts glyoxylate to glycine (Fig. 5b). Glycine can be converted by the glycine cleavage system (GCS) to 5,10-methylenetetrahydrofolate, a useful one-carbon pool for biosyntheses, and two key GCS proteins appear to have been lost independently from filaria and tapeworms, suggesting their GCS is non-functional (Supplementary Table 19e). In addition, filaria have lost the ability to produce and use ketone bodies, a temporary store of acetyl coenzyme A (CoA) under starvation conditions (Supplementary Table 19b). The filaria lost these features after they diverged from *D. medinensis*, an outgroup to the filaria in clade IIIc that has a major difference in its life cycle, namely, a free-living larval stage (Supplementary Table 12).

The absence of multiple initial steps of pyrimidine synthesis was observed in some nematodes, including all filaria (as previously reported\(^\text{27}\)) and topeworms, suggesting they obtain pyrimidines from *Wolbachia* endosymbionts or from their hosts, respectively (Supplementary Table 19f). Similarly, all platyhelminths and some nematodes (especially clade IVa and filaria IIIc) appear to lack key genes for purine synthesis (Supplementary Table 19g) and rely on salvage instead. However, despite the widespread belief that nematodes cannot synthesize purines\(^\text{90,91}\), complete or near-complete purine synthesis pathways were found in most members of clades I, IIIb and V. Nematodes are known to be unable to synthesize haem\(^\text{47}\), but the pathway was found in platyhelminths, including *S. mansoni* (despite conflicting biochemical data\(^\text{27}\)) (Supplementary Table 19h and Supplementary Table 20i).

Genes from the β-oxidation pathway, used to break down lipids as an energy source, were not detected in schistosomones and some cyclophyllidean tapeworms (*Hymenolepis, Echinococcus*: Fig. 5a, M00087; Supplementary Table 19a). These species live in glucose-rich environments and may have evolved to use glucose and glycogen as principal energy sources. However, biochemical data suggest they do perform β-oxidation\(^\text{92}\), so they may have highly diverged but functional β-oxidation genes.

The lactate dehydrogenase (LDH) pathway is a major source of ATP in anaerobic but glucose-rich environments. Platyhelminths have high numbers of LDH genes, as do blood-feeding *Ancylostoma* hookworms (Supplementary Fig. 22g). Nematode clades Vc (including *Ancylostoma*) and IIIb have expansions of α-glucoisidases that may break down starch and disaccharides in host food to glucose (Supplementary Fig. 22a). Many nematodes and flatworms use malate dismutation as an alternative pathway for anaerobic ATP production\(^\text{28}\). The importance of the pathway for clade IIIb nematodes was reflected in expanded families encoding two key pathway enzymes PEPC and methylmalonyl CoA epimerase, and the intracellular trafficking chaperone for cobalamin (vitamin B-12), a cofactor for the pathway (Supplementary Fig. 22c–e and Supplementary Table 9a). A second cobalamin-related family (CobQ/CbiP) is clade IIIb-specific and appears to have been gained by horizontal gene transfer from bacteria (Supplementary Fig. 23a, Supplementary Note 2.6 and Methods). A glutamate dehydrogenase family expanded in clade IIIb (Supplementary Fig. 22h) is consistent with a GABA (γ-aminobutyric acid) shunt that helps maintain redox balance during malate dismutation. In clade Va, an
expansion in the propionate breakdown pathway\textsuperscript{44} (Supplementary Fig. 22f), suggested degradation of propionate, originating from malate dismutation or fermentation in the host’s stomach\textsuperscript{35}. Clade I nematodes have an acetate/succinate transporter that appeared to have been gained from bacteria (Supplementary Note 2.6 and Methods), and may participate in acetate/succinate uptake or efflux (Supplementary Fig. 23b).

Identifying new anthelmintic drug targets and drugs. As an alternative to a purely target-based approach that would require extensive compound screening, we explored drug repurposing possibilities. We developed a pipeline to identify the most promising targets from parasitic nematodes and platyhelminths. These sequences were used in searches of the ChEMBL database that contains curated activity data on defined targets in other species and their associated drugs and compounds (Supplementary Note 5 and Methods). Our pipeline identified compounds that are predicted to interact with the top 15% of highest-scoring worm targets ($n=289$). These targets included 17 out of 19 known or likely targets for World Health Organization-listed anthelmintics that are represented in ChEMBL (Supplementary Table 21b). When compounds within a single chemical class were collapsed to one representative, this potential screening set contained 5,046 drug-like compounds, including 817 drugs with phase III or IV approval and 4,229 medicinal chemistry compounds (Supplementary Table 21d). We used a self-organizing map to cluster these compounds based on their molecular fingerprints (Fig. 6). This classification showed that the screening set was significantly more structurally diverse than existing anthelmintic compounds (Supplementary Fig. 24).

The 289 targets were further reduced to 40 high-priority targets, based on predicted selectivity, avoidance of side-effects (clade-specific chokepoints or lack of human homologues) and putative vulnerabilities, such as those suggested by gene family expansions in parasite lineages, or belonging to pathways containing known or likely anthelmintic targets (Supplementary Fig. 25). These 40 targets were associated with 720 drug-like compounds comprising 181 phase III/IV drugs and 539 medicinal chemistry compounds. There is independent evidence that some of these have anthelmintic activity. For example, we identified several compounds that potentially target glycogen phosphorylase, which is in the same pathway as a likely anthelmintic target (glycogen phosphorylase phosphatase, likely target of niridazole; Supplementary Fig. 25). These compounds included the phase III drug alvodicol (flavopiridol), which has anthelmintic activity against C. elegans\textsuperscript{36}. Another example is the target cathepsin B, expanded in nematode clade Va (Supplementary Table 9a), for which we identified several compounds including the phase III drug odanacatib, which has been shown to have anthelmintic activity against hookworms\textsuperscript{37}. Existing drugs such as these are attractive candidates for repurposing and fast-track therapy development, while the medicinal chemistry compounds provide a starting point for broader anthelmintic screening.

Discussion

The evolution of parasitism in nematodes and platyhelminths occurred independently, starting from different ancestral gene sets and physiologies. Despite this, common selective pressures of adaptation to host gut, blood or tissue environments, the need to avoid hosts’ immune systems, and the acquisition of complex life cycles to effect transmission, may have driven adaptations in common biological pathways. While previous comparative analyses of parasitic worms have been limited to a small number of species within narrow clades, we have surveyed parasitic worms spanning two phyla, with a focus on those infecting humans and livestock. A large body of draft genome data (both published and unpublished) was utilized but, by focusing on lineage-specific trends rather than individual species-specific differences, our analysis was deliberately conservative. In particular, we have focused on large gene family expansions, supported by the best-quality data and for which functional information was available. Sequencing of further free-living species, better functional characterization, and identification of remote orthologs (particularly for platyhelminths\textsuperscript{87}), will undoubtedly refine the resolution of parasite-specific differences, but our gene family analyses have already revealed expansions and synein morphologies in functional classes of likely importance to parasitism, such as feeding and interaction with hosts. We have used a drug repurposing approach to predict potential new anthelmintic drug targets and drugs/drug-like compounds that we urge the community to explore. Further new potential drug targets, worthy of high-throughput compound screening, may be exposed by the losses of key metabolic pathways and horizontally acquired genes that we find in particular parasite groups. This is an unprecedented dataset of parasitic worm genomes that provides a new type of pan-species reference and a much needed stimulus to the study of parasitic worm biology.

URLs. SMALT, http://www.sanger.ac.uk/science/tools/smalt-0; RepeatModeler, http://www.repeatmasker.org/RepeatModeler.html; TransposonPSI, http://transposonpsi.sourceforge.net; RepeatMasker, http://www.repeatmasker.org; code for calculating gene family metrics, http://tinyurl.com/comparaFamiliesAnalysis.py; WormBase ParaSite, https://parasite.wormbase.org/.
58. Esteban, M. R., Giovinazzo, G., de la Hera, A. & Goday, C. PUMA1: a novel protein that associates with the centrosomes, spindle and centromeres in the nematode Caenorhabditis elegans. J. Cell. Sci. 11, 723–735 (1994).

59. Tobihi, H., Ettorre, A. & Muller, F. Chromatin dimunition in nematode development. Trends Genet. 8, 427–-432 (1992).

60. Albarqui, M. M. et al. Regulation of life cycle checkpoints and developmental activation of infective larvae in Strongyloides stercoralis by dafachronic acid. PLoS Pathog. 12, e1005358 (2016).

61. Zarlenga, D. S., Nabet, A. J., Gasbarre, L. C. & Garrett, W. M. A calcium-activated nucleotidase secreted from Osetaria actegagi 4-stage larvae is a member of the novel salivary alyposes present in blood-feeding arthropods. Parasitology 138, 333–343 (2011).

62. Catthcart, M. K. & Bhattacharjee, A. Monoamine oxidase A (MAO-A): a signature marker of alternatively activated monocytes/macrophages. Inflamm. Cell Signal. e1, e161 (2014).

63. Coxley, G., Maizels, R. M. & Buck, A. H. Eosomes and other extracellular vesicles: the new communicators in parasite infections. Trends Parasitol. 31, 477–489 (2015).

64. Wu, C. et al. Mapping the binding between the tetraspanin molecule (Sc23) of Schistosma mansoni and human non-immune IgG. PLoS ONE 6, e19112 (2011).

65. Krautz-Peterson, G. et al. Schistosoma mansoni infection of mice, rats and humans elicits a strong antibody response to a limited number of reduction-sensitive epitopes on five major tegumental membrane proteins. PLoS Negl. Trop. Dis. 11, e0005306 (2017).

66. Prior, A. et al. A surface-associated retinol- and fatty acid-binding protein (Gp-FAR-1) from the potato cyst nematode Globodera pallida: lipid binding activities, structural analysis and expression pattern. Biochem. J. 356, 387–394 (2001).

67. Ray-Burisco, M. F. et al. Diversity in the structures and lignin-binding sites of nematode fatty acids and retinol-binding proteins revealed by Na-FAR-1 from Neutor americana. Biochem. J. 471, 403–414 (2015).

68. Dell, A., Haslam, S. M. & Morris, H. R. in Parasitic Nematodes: Molecular Biology, Biochemistry and Immunology (eds. Kennedy, M. W. & Harnett, W.) 285–307 (Cabi Publishing, Oxfordshire, UK, 2013).

69. Rodrigues, J. A. et al. Parasite glycochemistry: a bittersweet symphony. Cell Microbiol. 11, e1005169 (2015).

70. Anderson, L. et al. Schistosoma mansoni egg, adult male and female comparative gene expression analysis and identification of novel genes by RNA-Seq. PLoS Negl. Trop. Dis. 9, e0004334 (2015).

71. Gong, H. et al. A novel PAN/apple domain-containing protein from Toxoplasma gondii: characterization and receptor identification. PLoS ONE 7, e30169 (2012).

72. 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 (2009).

73. McKiernon, J. H., Caffrey, C., Kelly, B., Loke, P. & Sajid, M. Proteases in parasitic diseases. Annu. Rev. Pathol. 1, 497–536 (2006).

74. Williamson, A. L. et al. Ancylostoma caninum MTP-1, an astacin-like metalloproteinase secreted from infective hookworm larvae, is involved in tissue migration. Infect. Immun. 74, 961–967 (2006).

75. Williamson, A. L. et al. A multi-enzyme cascade of hemoglobin proteolysis in the intestine of blood-feeding hookworms. J. Biol. Chem. 279, 35950–35957 (2004).

76. Delcroix, M. et al. A multienzyme network functions in intestinal protein digestion by a phyletminal parasite. J. Biol. Chem. 281, 39316–39329 (2006).

77. Duffy, M. S., Cavaco, D. K., Zarlenga, D. S., Sukhumavasi, W. & Appleton, J. A. Cathepsin B homologue at the interface between a parasitic nematode and its intermediate host. Infect. Immun. 74, 1297–1304 (2006).

78. Cancela, M. et al. A distinctive repertoire of cathepsins is expressed by juvenile invasive Fasciola hepatica. Biochem. 98, 1461–1475 (2008).

79. kroon, D. P. Proteinase inhibitors and helminth parasite infection. Parasite Immunol. 29, 57–71 (2007).

80. Rehman, A. A., Ahsan, H. & Khan, F. H. Alpha-2-macroglobulin: a multifunctional proteinase inhibitor, kills hookworm in vivo. Parasit. 10, 313 (2014).

81. MacDonald, K. et al. Functional characterization of a novel family of acetylcholine-gated chloride channels in Schistosoma mansoni. PLoS Pathog. 10, e1004181 (2014).

82. Courtot, E. et al. Functional characterization of a novel class of morantel-sensitive acetylcholine receptors in nematodes. PLoS Pathog. 11, e1005267 (2015).

83. Van Soest, P. J. Nutritional Ecology of the Ruminant (Cornell Univ. Press, Ithaca, NY, USA, 1994).

84. Taylor, C. M. et al. Using existing drugs as leads for broad spectrum anthelmintics targeting protein kinases. PLoS Pathog. 9, e1003149 (2013).

85. Vermeire, J. J., Suzuki, B. M. & Caffrey, C. R. Odanacatib, a cathepsin K cysteine protease inhibitor, kills hookworm in vivo. Pharmaceut. 39, 16 (2016).

Acknowledgements

We thank the WSI DNA Pipeline teams, particularly C. Griffiths, N. Park, L. Shirley, M. Quail, D. Willey and M. Jones; WSI Pathogen Informatics, especially J. Keane; T.D. Otto for bioinformatics advice; MGI faculty and staff, especially M. Schmidt, C. Fronick, N. Cordes, T. Miner, R. Fulton and other members of the Project Management, Resource Bank, Library Construction and Data Production teams; D. Hughes, M. Muftulito at the European Bioinformatics Institute, for support running Maker and Ensemble Compar; and K. Gharbi and his staff at Edinburgh Genomics for support; V. Gelmedin, R. Fujisawa, F. Brael, the late Purnomo (University of Indonesia, Jakarta), J. Ahrens, E. Hernández Redondo, F. Jackson, E. Redman, A. Ito, J. Saldaña, M. Fernanda Dominguez, W. Gause, M. Badets, I.E. Samonte, A. Koehler, M. Nielsen, L.S. Mansfield, T. Sonstegard for sample preparation. The work was supported by funding from Wellcome (206194), Medical Research Council (MR/L001020/1), and Biotechnology and Biological Sciences Research Council (BB/K020048/1) to M.B., and US National Institutes of Health (NIH)—National Human Genome Research Institute grant number U54HG003079, National Institute of Allergy and Infectious Diseases grant number AI081803, and National Institute of General Medical Sciences grant number GM097435 to M.G. Genomic sequencing and analysis in Edinburgh was supported by EU ST2A award 242131 ‘Enhanced Protective Immunity Against Filariasis’ (EPIAF) (to D.W.T.). S.A. Babayan was also supported by the EU project EPIAF. G. Koutsovoulos was supported by a BBiT/Exeter University PhD scholarship and D.R.L. by a joint Edinburgh University/James Hutton Institute PhD scholarship. J.E.A. was supported by MRC grant MR/K012077/1. J.P. and S.S. were supported by the National Institutes of Health/NIAID (R21 AI126466) and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2014-06664). Additional computing resources were provided through Compute Canada by the University of Toronto SciNet HPC Consortium. R.M.M. was supported by a Wellcome Investigator Award (ref. 106122) and Wellcome core funding to the Wellcome Centre for Molecular Parasitology (Ref 104111). J.B.M. was supported by the Scottish Government RESAS. T.S. was supported by the Institute of Parasitology, BC CAS (RVO: 60073443). A.R.L. and P.M. were supported by a Strategic Award from Wellcome (WT104104/Z/14/Z) and the Member States of the European Molecular Biology Laboratory (EMBL). Schistosome samples were obtained from the SCAN repository (Wellcome grant 104958).

Author contributions

Project leadership and conception: M.B. and M.R. Writing the manuscript: M.B., M.R., A.C., J.A.C., N.H., D.R.L., R.M.M. and B.Tyagi. Project planning or management: M.B., M.R., J.A.C., N.H., M.M. and A.G. Genome sequencing: H.B., H.P., M. Whitehead, J.H., R.M. PO, W.D.T. and A.T. and Z.X. Preparation or provision of parasite material or nucleic acid: F.A., J.E.A., K.A., S.A. Bisset, G.B., H.M.B., S.A. Babayan, T.C.B., E.C., J.C., P.J.C., C.C., E.D., M.L.E., A.E., K.S.P., F.S.G., Y.H., J.M.H., D.E.H., J.H., F.H., M.K., T.K., R.M.M., B.G., J.M., P.D.O., A.O., F.P., K.P., D.R., M.G.S., H.S., M. Schneyder, T.S.,...
V.N.T., D.W.T., R. Toledo, J.F.U., L.C.W. and D.Z. Production bioinformatics: A.C., J.A.C., D.G., B.H., J.L., P.O., D.M.R., B.A.R., E.S. and A.T. Annotation of genomes: A.C., B.H., K.L.H., G.Kaur, G.Koutsovoulos, S.K., D.R.L., J.M., P.O., K.H.P., B.A.R., E.S. and Z.X. Assembly of genomes at WSI: J.T.T. Assembly of genomes at MGI: J.M., P.O., K.H.P. and Z.X. Assembly of genomes at BaNG: G.Kaur, G.Koutsovoulos, S.K. and D.R.L. Assembly and annotation of mitochondrial genomes: T.K. Analysis of variation in genome size: A.C. and J.A.C. Analysis of GO terms and domains: A.C., D.R.L., J.L., A.I.R., B.A.R. and M.Shafie. Development of visualisation software and production of the species tree: J.A.C. Analysis of synapomorphies: M.L.B. and D.R.L. Analysis of expanded gene families: M.B., A.C., J.A.C., S.R.D., N.H., A.I.R., D.M.R., G.R., B.A.R. and R.Tyagi. Analysis of hypothetical genes: A.C. and A.J.R. Analysis of SCP/TAPS: H.M.K., T.J.K., T.J.L. and I.J.T. Analysis of proteases: N.D.R. Analysis of GPCRs: T.A.D., N.J.W. and M.Z. Analysis of ion channels: R.N.B. and M.Z. Analysis of kinases: J.M. and B.A.R. Analysis of metabolic pathways: A.C., J.A.C., T.K., J.S. and R.Tyagi. Chemo genomics analyses: A.C., J.A.C., A.R.L., J.L., P.M. and N.M.O.’B.

Competing interests
The authors declare no competing interests.

International Helminth Genomes Consortium

Avril Coghlan*1, Rahul Tyagi2, James A. Cotton1, Nancy Holroyd1, Bruce A. Rosa2, Isheng Jason Tsai3,4,5, Dominik R. Laetsch6, Robin N. Beech7, Tim A. Day8,9, Kymberlie Hallsworth-Pepin7, Huei-Mien Ke3, Tzu-Hao Kuo3, Tracy J. Lee3,4,5, John Martin2, Rick M. Maizels10, Prudence Mutowo11, Philip Ozersky2, John Parkinson12,13, Adam J. Reid1, Neil D. Rawlings11, Diogo M. Ribeiro14, Lakshmipuram Seshadri Swapna12,13, Eleanor Stanley1, David W. Taylor15, Nicolas J. Wheeler8,16, Mostafa Zamanian16, Xu Zhang2, Fiona Allan17, Judith E. Allen18, Kazuhiro Asano19, Simon A. Babayan20,21, Germanus Bah22, Helen Beasley1, Hayley M. Bennett1, Stewart A. Bisset23, Estela Castillo24, Joseph Cook25, Philip J. Cooper26,27, Teresa Cruz-Bustos28, Carmen Cuéllar29, Eileen Devaney20, Stephen R. Doyle1, Mark L. Eberhard30,31, Aidan Emery32, Keeseon S. Eom32, John S. Gilliard33, Daria Gordon1, Yvonne Harcus21,34, Bhavana Harsha1, John M. Hawdon35, Dolores E. Hill36, Jane Hodgkinson37, Petr Horák38, Kevin L. Howe39, Thomas Huckvale1, Martin Kalbe39, Gaganjot Kaur40, Taisei Kikuchi41, Georgios Koutsovoulos6,42, Sujaik Kumar6, Andrew R. Leach31, Jane Lomax1, Benjamin Makepeace37, Jacqueline B. Matthews43, Antonio Muro44, Noel Michael O’Boyle45, Peter D. Olson17, Antonio Osuna28, Felix Partono46, Kenneth Pfarr47, Gabriel Rinaldi1, Pilar Foronda48, David Rollinson17, Mercedes Gomez Samblas28, Hiroshi Sató49, Manuela Schnyder50, Tomáš Scholz51, Myriam Shafie1, Vincent N. Tanya22, Rafael Toledo52, Alan Tracey1, Joseph F. Urban36, Lian-Chen Wang53, Dante Zar lenga36, Mark L. Blaxter5, Makedonna Mitрева54* and Matthew Berriman*1

1Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. 2McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA. 3Biodiversity Research Center, Academia Sinica, Taipei, Taiwan. 4Department of Life Science, National Taiwan Normal University, Taipei, Taiwan. 5Biodiversity Program, Academia Sinica and National Taiwan Normal University, Taipei, Taiwan. 6BaNG, Institute of Evolutionary Biology and Centre for Immunity, Infection and Evolution, University of Edinburgh, Edinburgh, UK. 7Institute of Parasitology, McGill University, Montreal, Quebec, Canada. 8Department of Biomedical Sciences, Iowa State University, Ames, IA, USA. 9Institute for Food Security, Queen’s University Belfast, Belfast, UK. 10Wellcome Centre for Molecular Parasitology, University of Glasgow, Glasgow, UK. 11European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK. 12Program in Molecular Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada. 13Departments of Biochemistry and Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. 14Aix-Marseille Université, Marseille, France. 15School of Biological Sciences, University of Edinburgh, Edinburgh, UK. 16Department of Pathobiological Sciences, University of Wisconsin-Madison, Madison, WI, USA. 17Natural History Museum, London, UK. 18Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. 19Department of Physiology, Showa University, Tokyo, Japan. 20Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow, UK. 21Institute of Immunology and Infection Research, University of Edinburgh, Edinburgh, UK. 22Institut de Recherche Agricole pour le Développement, Ngao undéré, Cameroon. 23Hopkirk Research Institute, AgResearch Ltd, Palmerston North, New Zealand. 24Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay. 25Museum of Southwestern Biology, University of New Mexico, Albuquerque, NM, USA. 26Institute of Infection and Immunity, St. George’s, University of London, London, UK. 27Facultad de Ciencias Medicas, de la Salud y la Vida, Universisdad Internacional del Ecuador, Quito, Ecuador. 28Department of Parasitology, University of Granada, Granada, Spain. 29Departamento de Parasitología, Universidad Complutense de Madrid, Madrid, Spain. 30Division of Parasitic Diseases and Malaria, Centers for Disease Control and Prevention, Atlanta, GA, USA. 31The Carter Center, Atlanta, GA, USA. 32Department of Parasitology, Chungbuk National University School of Medicine, Cheongju, Korea. 33Department of Comparative Biology and Experimental Medicine, University of Manchester, Manchester, UK.
Articles

University of Calgary, Calgary, Alberta, Canada. 34Centre for Cardiovascular Science, Queen’s Medical Research Institute, Edinburgh, UK. 35Department of Microbiology, Immunology, and Tropical Medicine, The George Washington University, Washington, DC, USA. 36United States Department of Agriculture, Beltsville Agricultural Research Centre, Beltsville, MD, USA. 37Department of Infection Biology, University of Liverpool, Liverpool, UK. 38Department of Parasitology, Charles University, Prague, Czech Republic. 39Department of Evolutionary Ecology, Max Planck Institute for Evolutionary Biology, Ploen, Germany. 40The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada. 41Faculty of Medicine, University of Miyazaki, Miyazaki, Japan. 42INRA PACA Site de Sophia-Antipolis, Sophia Antipolis, France. 43Moredun Research Institute, Edinburgh, UK. 44Biomedical Research Institute of Salamanca-Research Centre for Tropical Diseases at the University of Salamanca (IBSAL-CIETUS), University of Salamanca, Salamanca, Spain. 45NextMove Software Ltd, Cambridge, UK. 46Department of Parasitology, University of Indonesia, Jakarta, Indonesia. 47Institute for Medical Microbiology, Immunology and Parasitology, University of Bonn Medical Center, Bonn, Germany. 48Institute of Tropical Diseases and Public Health of the Canary Islands, Universidad de La Laguna, Tenerife, Spain. 49Joint Faculty of Veterinary Medicine, Yamaguchi University, Yamaguchi, Japan. 50Institute of Parasitology, University of Zurich, Zurich, Switzerland. 51Biology Centre CAS, Institute of Parasitology, Branišovská, Czech Republic. 52Departamento de Parasitología, Universidad de Valencia, Valencia, Spain. 53Department of Parasitology, Chang Gung University, Taoyuan, Taiwan. 54Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO, USA. *e-mail: mb4@sanger.ac.uk; alc@sanger.ac.uk; mmitreva@wustl.edu
Methods
Sample collection and preparation. Sources of material and sequencing approaches are summarized in Supplementary Table 1.

Wellcome Sanger Institute (WSI) data production. The genomes of 36 species (Supplementary Tables 1 and 2) were sequenced at WSI. The C. elegans N2 was also resequenced at WSI.

WSI sequencing and assembly. PCR-free 400–550 bp paired-end Illumina libraries were prepared from ~0.8 ng to ~5 ng genomic DNA, as described for Strongyloides stercoralis18. Where there was insufficient DNA, adapter-ligated material was subjected to ~8 PCR cycles.

We used 1–10 µg gDNA or whole genome amplification DNA to generate 3 kb mate-pair libraries, as described for S. stercoralis18. If there was insufficient gDNA, whole genome amplification was performed using GenomiPhi v2. Each library was run on ≥2 Illumina HiSeq 2000 lanes.

Short insert paired-end reads were corrected and assembled with SGA v0.9.7 (Supplementary Fig. 26a). This assembly was used to calculate the k-mer distribution for all odd k of 41–81, using GenomeTools v1.3.799. The k-mer length for which the maximum number of unique k-mers was present was used as the k-mer setting in a second assembly, using Velvet v1.2.03187 with SGA-corrected reads. For species with 3 kb mate-pair data, the Velvet assembly was scaffolded using SPSPACE188. Contigs were extended, and gaps closed and shortened, using Gapfiller189 and IMAGE190. Short fragment reads were remapped to the assembly using SMALT (see URLs), and unaligned reads assembled using Velvet187 and this merged with the main assembly. The assembly was re-scaffolded using SSPACE188, and consensus base quality improved with iCOR191, REAP192 was used to break incorrectly assembled scaffolds/contigs. We carried out manual improvement for Wuchereria bancrofti and D. medinensis using Gapgs193 and Illumina read-pairs.

WSI assembly quality control. Contamination screening. Assemblies were screened for contamination using BLAST− against vertebrate and invertebrate sequences (see ref. 194). For Anisakis simplex, the assembly contained minor laboratory contamination with S. mansoni, which we removed using BLASTN against S. mansoni.

Assembly completeness. CEGMA v2.4109 was used to assess completeness. Consistent sets of CEGMA genes were missing from some phylogenetic groups (Supplementary Table 2); these were discounted from the completeness calculation for those species (CEGMA in Supplementary Table 2).

Effect of repeats. We re-mapped the short-insert library’s reads to the appropriate assembly using SMALT (see URLs; indexing k = 23 - 43 and mapping -y 0.9 - x - r 1). For each scaffold of ≥8 kb, median (med) and mean (m) per-base read-depth were calculated using BEDTools195, and genome-wide depth (med) calculated as the median med, (ref. 196). For a scaffold, the extra sequence that would be gained by ‘uncollapsing’ repeats was estimated as e = (m - med) x l/s med, (Supplementary Table 5).

WGS gene prediction. Our pipeline197 had four steps (Supplementary Fig. 27a). First, repeats were masked. Second, preliminary gene predictions, to use as input for MAKER, were generated using AUGUSTUS113 (- b 4 - h - b 5 = b) and its likelihood calculated on a maximum-parsimony guide tree for all relatively simple (single-matrix) amino acid substitution models in Mafft v6.857198 and used as the input of MAKER (see Methods) to create a mixed-effect model, the species tree (see Methods) was transformed into an ultrametric tree using PATHd8199, with a small constant added to short branches to ensure no zero-length branches were reconstructed; and outgroup species were removed.

Compara database. An in-house Ensembl Compara database was constructed containing the 81 platyhelminths and nematodes, and 10 additional outgroups (Supplementary Table 2). All parasitic nematode/platyhelminth species with gene sets available at the time (April 2014) were included.

The species tree used to construct the initial version of our database was used an edited version of the National Center for Biotechnology Information (NCBI) taxonomy200 with several controversial speciation nodes represented as multifurcations. For our final database, the input species tree was derived by building a tree based on the previous database version, based on one-to-one ortholog presence in ≥20 species. To do this, proteins in each ortholog group were aligned using MAFFT v6.875201, alignment trimming using GBAlign v0.918202, concatenated and used to build a maximum likelihood tree using a partitioned analysis in RAXML v7.8.6203, using the minimum Akaike’s information criterion (minAIC) model for each ortholog group.

The database was queried to identify gene families, orthologs and paralogs.

Species tree and tree based on gene family presence. We identified 202 gene families present in ≥25% of the 81 helminths and 10 outgroups in our Compara database (Methods) and always single-copy. For each family, amino acid sequences were aligned using MAFFT v7.205204 (-auto). Each alignment was trimmed using GBAlign v0.918202, and its likelihood calculated on a maximum-parsimony guide tree for all relatively simple (single-matrix) amino acid substitution models in Mafft v8.0.24205, and the minAIC model identified. Alignments were concatenated and a maximum-likelihood tree built, under a partitioned model in which sites from a gene were assigned the minAIC model for that gene, with a discrete gamma distribution of rates across sites. Relationships within outgroup lineages were constrained to match the standard view of metazoan relationships (for example, Dunn et al.199). The final tree was the highest likelihood one from five search replicates with different random number seeds. One hundred bootstrap resampling replicates were performed, each based on a single rapid search.

We also constructed a maximum-likelihood phylogeny based on gene family presence/absence for families not shared by all 81 nematode/platyhelminth species, using RaxML v8.2.8206, with a two-state model and the Lewis method to correct for absence of constant-state observations.

Functional annotation. InterProScan207 v5.0.7 was used to identify conserved domains from all predicted proteins. A name was assigned to each predicted protein based on curated information in UniProt208 for orthologs identified from our Compara database (Methods), or based on InterPro209 domains (see ref. 210). Gene ontology (GO) terms were assigned by transferring GO terms from orthologs210, and using InterProScan.

Transcriptome sequencing and gene prediction. Transcriptome libraries (Supplementary Table 22) were generated with the Illumina TS stranded protocol, and reads assembled using Trinity211 (see ref. 212). Genes were predicted using MAKER212, based on input gene models from SNAP213, FGENESH (Softberry), Augustus214, and aligned messenger RNA, EST, transcriptome and protein data from the same or related species (Supplementary Fig. 27b; see ref. 215).
Signal peptides and transmembrane domains were predicted using Phobius\(^{10}\) v.0.1 and SecretomeP\(^{10}\) v.1.0. A protein predicted by Phobius to have a transmembrane domain was categorized as ‘membrane-bound,’ and non-membrane-bound proteins as ‘classically secreted.’ If Phobius predicted a signal peptide within 70 amino acids of their start. Remaining proteins in which SecretomeP predicted a signal peptide were classified as ‘non-classically secreted’ (Supplementary Table 7).

Pairwise combinations of Pfam domains were identified in proteins of the 81 nematodes and platyhelminths. After excluding those present in complete genomes of other phyla in UniProt (June 2016), we classified a combination as ‘nematode-specific’ (or ‘flatworm-specific’) if it was present in >30% of nematodes (platyhelminths) and no platyhelminths (nematodes) (Supplementary Table 14).

Synapomorphic gene families. Families in our Compara database (Methods) were extracted by screening for a Ferrochelatase (IPR001015) with a particular InterPro domain. Counts of proteins annotated in each particular species outside the node of interest, and (2) in which ≥1 gene of a particular species was present in the node of interest, and must not contain other species.

Potential drug targets. Nematode and platyhelminth proteins from tier 1 species (with high-quality assemblies; Methods) were searched against single-protein targets from CheMBL\(^{21}\) using BLASTP (E ≤ 10\(^{-10}\)). After collapsing by gene family, 1,925 worm genes remained. To assign a ‘target score’ to each worm gene, the main factors considered were similarity to known drug targets; lack of human homologues; and whether C. elegans/Drosophila melanogaster homologues had lethal phenotypes (see ref. 19).

Potential new anthelmintic drugs. CheMBL\(^{21}\) was used to identify 827,889 compounds with activities against CheMBL targets to which worm proteins had BLAST matches. To calculate ‘compound scores,’ we prioritized compounds in high clinical development phases, oral/topical administration, crystal structures, properties consistent with oral drugs and lacking toxicity (see ref. 19).

A ‘diverse screening set.’ The 131,452 candidates were placed into 27,944 chemical classes, based on ECFP4 fingerprints (see ref. 20). They were filtered by (1) discarding medicinal chemistry compounds that did not co-appear in a PDBe structure with the CheMBL target; or (2) had median pChEMBL > 5; leading 131,452 ‘top drug candidates.’
grid. The self-organizing map was trained for 4,000 steps, where training optimized Tanimoto distances between ECFP4 fingerprints.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

Sequence data have been deposited in the European Nucleotide Archive (ENA). Assemblies and annotation are available at WormBase and WormBase-ParaSite (https://parasite.wormbase.org/). All have been submitted to GenBank under the BioProject IDs listed in Supplementary Table 1.

**References**

98. Simpson, J. T. & Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 22, 549–556 (2012).

99. Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans. Comput. Biol. Bioinform. 10, 645–656 (2013).

100. Zerbino, D. R. & Birney, E. Velvet: a novel approach for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–828 (2008).

101. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembly contigs using SSPACE. Bioinformatics 27, 578–579 (2011).

102. Butler, J. M. & Pirovano, W. Toward almost closed genomes with GapFiller. Genome Biol. 13, R56 (2012).

103. Tsai, I. J., Otto, T. D. & Bermiran, M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol. 11, R41 (2010).

104. Otto, T. D., Sanders, M., Bermiran, M. & Newbold, C. Iterative correction of reference nucleotides (iCORN) using second generation sequencing technology. Bioinformatics 26, 1704–1707 (2010).

105. Hunt, M. et al. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 14, R47 (2013).

106. Bonfield, J. K. & Whitwham, A. Gap5—editing the billion fragment parasitic worm genome assemblies. Nucleic Acids Res. 47, 11.12.1–11.12.34 (2014).

107. Allschl, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402 (1997).

108. Coghlan, A. L., Gordon, D. & Bermiran, M. Contamination screening of parasitic worm genome assemblies. PLoS. Biol. https://doi.org/10.1371/journal.pbio.0060192 (2008).

109. Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. Nucleic Acids Res. 37, 289–297 (2009).

110. Quinlan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. Curr. Protoc. Bioinformatics 47, 11.12.1–11.12.34 (2014).

111. Stanley, E., Coghlan, A. L. & Bermiran, M. A MAKER pipeline for prediction of protein-coding genes in parasitic worm genomes. PLoS. ONE 3, e2856 (2008).

112. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome software library for efficient processing of structured genome annotations. Genome Res. 20, 549–556 (2012).

113. Shibahara, N. et al. gnell: using BLAST searches to build homologous gene models. Bioinformatics 27, 2141–2143 (2011).

114. Otto, T. D., Dillon, G. P., De Grave, W. S. & Bermiran, M. RATT: rapid assembly of parasitic worm genome assemblies. Nucleic Acids Res. 41, 509–518 (2013).

115. Cochrane, G., Yandell, M., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for advanced protein alignment trimming. Bioinformatics 27, 2471–2477 (2011).

116. Berriman, M. et al. The fluka genome of filarial nematodes supports both metabolic and defensive roles within the symbiosis. Genome Res. 22, 2467–2477 (2012).

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

118. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for advanced protein alignment trimming. Bioinformatics 27, 2471–2477 (2011).

119. Wheeler, N., Day, T., Zamanian, M. & Kimber, M. GPCR identification in parasitic worm genome assemblies. Nucleic Acids Res. 40, D674–D682 (2012).

120. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. Mol. Biol. Evol. 28, 1928–1930 (2011).

121. Dobson, S. C., Yandell, M., Berriman, M. & Bermiran, M. Identification of lineage-specific gene family expansions in a database of gene families. PLoS. ONE 10, e0128750 (2015).

122. Kumar, S. & Blaxter, M. L. Simultaneous genome sequencing of symbionts and their hosts. Symbiosis 55, 119–126 (2011).

123. Crisoe, M. R. et al. The kmer software package: enabling efficient nucleotide sequence analysis. F1000Res. 6, 900 (2015).

124. Simpson, I. T. et al. ABYSS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123 (2009).

125. Darby, A. C. et al. Analysis of gene expression from the Wolbachia genome of a filarial nematode supports both metabolic and defensive roles within the symbiosis. Genome Res. 22, 2467–2477 (2012).

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

127. Ellingshaus, D., Kurtz, S. & Willhoef, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9, 18 (2008).

128. Jones, A. K., Davis, P., Hodgkin, J. & Sattelle, D. B. The nicotinic acetylcholine receptor gene family of the nematode Caenorhabditis elegans: an update on nomenclature. Invert. Neurosci. 7, 129–131 (2007).
158. Li, B. W., Rush, A. C. & Weil, G. J. Expression of five acetylcholine receptor subunit genes in 
Brugia malayi adult worms. Int. J. Parasitol. Drugs Drug Resist. 5, 100–109 (2015).

159. Buxton, S. K. et al. Investigation of acetylcholine receptor diversity in a 
nematode parasite leads to characterization of tribendimidine- and 
derquantel-sensitive nAChRs. PLoS Pathog. 10, e1003870 (2014).

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

161. Tusnady, G. E. & Simon, I. The HMMTOP transmembrane topology 
prediction server. Bioinformatics 17, 849–850 (2001).

162. Ronquist, F. et al. MrBayes 3.2: efficient Bayesian phylogenetic inference 
and model choice across a large model space. Syst. Biol. 61, 539–542 (2012).

163. Miranda-Saavedra, D. & Barton, G. J. Classification and functional 
anotation of eukaryotic protein kinases. Proteins 68, 893–914 (2007).

164. Martin, J. & Mitreva, M. Kinase annotation for helminths. 
Protoc. Exch. https://doi.org/10.1038/protex.2018.042 (2018).

165. Hung, S. S., Wasmuth, J., Sanford, C. & Parkinson, J. DETECT—a density 
estimation tool for enzyme classification and its application to Plasmodium 
falciparum. Bioinformatics 26, 1690–1698 (2010).

166. Claudel-Renard, C., Chevalet, C., Faraut, T. & Kahn, D. Enzyme-specific 
profiles for genome annotation: PRIAM. Nucleic Acids Res. 31, 
6633–6639 (2003).

167. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an 
automatic genome annotation and pathway reconstruction server. Nucleic 
Acids Res. 35, W182–W185 (2007).

168. Chang, A. et al. BREND A in 2015: exciting developments in its 25th year of 
existence. Nucleic Acids Res. 43, D439–D446 (2015).

169. Swapna, S., Tyagi, R., Mitreva, M. & Parkinson, J. Annotating metabolic 
enzymes in parasitic worm proteomes. Protoc. Exch. https://doi.org/10.1038/ 
protex.2018.047 (2018).

170. Karp, P. D. et al. Pathway Tools version 19.0 update: software for 
pathway/genome informatics and systems biology. Brief. Bioinformatics 17, 
877–890 (2016).

171. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes 
and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 
40, D742–D753 (2012).

172. Lombard, V., Golaoncda Ramulu, H., Drula, E., Coutinho, P. M. & 
Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. 
Nucleic Acids Res. 42, D490–D495 (2014).

173. Tyagi, R., Swapna, S., Parkinson, J. & Mitreva, M. Comparative analysis of 
metabolism in parasitic worms. Protoc. Exch. https://doi.org/10.1038/ 
protex.2018.048 (2018).

174. Lombard, V., Golaoncda Ramulu, H., Drula, E., Coutinho, P. M. & 
Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. 
Nucleic Acids Res. 42, D490–D495 (2014).

175. Real, R. & Vargas, J. M. The probabilistic basis of Jaccard's index of 
similarity. Syst. Biol. 45, 380–385 (1996).

176. Taylor, C. M. et al. Discovery of anthelmintic drug targets and drugs 
using chokepoints in nematode metabolic pathways. PLoS Pathog. 9, 
e1003505 (2013).

177. Gaulton, A. et al. The ChEMBL database in 2017. Nucleic Acids Res. 45, 
D945–D954 (2017).

178. Coghlan, A. L. et al. Creating a screening set of potential anthelmintic 
compounds using ChEMBL. Protoc. Exch. https://doi.org/10.1038/ 
protex.2018.053 (2018).

179. Velankar, S. et al. PDBe: improved accessibility of macromolecular structure 
data from PDB and EMDB. Nucleic Acids Res. 44, D385–D395 (2016).

180. Sterling, T. & Irwin, J. J. ZINC 15—ligand discovery for everyone. J. Chem. 
Inf. Model. 55, 2324–2337 (2015).

181. Wehrens, R. & Buydens, L. M. C. Self- and super-organizing maps in R: the 
kohonen package. J. Stat. Softw. 21, 1–19 (2007).
Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

| n/a | Confirmed |
|-----|-----------|
| ☒   | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☒   | An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒   | The statistical test(s) used AND whether they are one- or two-sided |
| ☒   | Only common tests should be described solely by name; describe more complex techniques in the Methods section. |
| ☒   | A description of all covariates tested |
| ☒   | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☒   | A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒   | For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted. Give P values as exact values whenever suitable. |
| ☒   | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒   | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒   | Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated |
| ☒   | Clearly defined error bars |
| ☒   | State explicitly what error bars represent (e.g. SD, SE, CI) |

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

| Data collection | No software was used to collect the data in the study. |
|-----------------|------------------------------------------------------|
| Data analysis   | A large number of software applications were used in this study. All software used (custom and commercial/publicly available) are listed in the Methods. All custom scripts are available on request from the corresponding authors. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequence data have been deposited in the European Nucleotide Archive (ENA). Assemblies and annotation are available at WormBase and WormBase-ParaSite. All have been submitted to GenBank under BioProjects listed in Supplementary Table 1.
Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
Sample size was determined by the availability of parasite material. All samples were surplus material from other ongoing research projects, and due to the difficulties involved with obtaining parasite material, sample size was determined primarily by sample availability, rather than a predetermined number.

Data exclusions
Some samples provided for this study were of poor quality, and thus the resulting data was of insufficient quality to warrant inclusion in the data set. Exclusion criteria were not predetermined.

Replication
Experimental findings were not reproduced due to the scale of the study, in terms of time and cost, combined with the issue associated with obtaining parasite material.

Randomization
Allocation of samples into experimental groups was done so based on taxonomic classification.

Blinding
Blinding was not relevant to this study as analysis were explicitly comparative.

Reporting for specific materials, systems and methods

Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
|     | Unique biological materials |
|     | Antibodies |
|     | Eukaryotic cell lines |
|     | Palaeontology |
|     | Animals and other organisms |
|     | Human research participants |

Methods

| n/a | Involved in the study |
|-----|-----------------------|
|     | ChIP-seq |
|     | Flow cytometry |
|     | MRI-based neuroimaging |

Unique biological materials

Policy information about availability of materials

Obtaining unique materials
Not all unique materials used in this study are available due to a number of them being from wild or livestock animals, rather than laboratory maintained animals. They are either unique samples that could not easily be obtained again, or all the available sample has been used up in this experiment. In some cases, material from laboratory maintained populations may be available on request where feasible.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals
Samples obtained were parasite materials that were surplus to other existing ongoing projects, either from wild animals, laboratory animals or already dead animals (e.g. from an abattoir). Further details on the samples are given in Supplementary Table 1.

Wild animals
Samples obtained were parasite materials that were surplus to other existing ongoing projects, either from wild animals, laboratory animals or already dead animals (e.g. from an abattoir). Further details on the samples are given in Supplementary Table 1.

Field-collected samples
Samples obtained were parasite materials that were surplus to other existing ongoing projects, either from wild animals,
Field-collected samples include laboratory animals or already dead animals (e.g. from an abattoir). Further details on the samples are given in Supplementary Table 1.