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 through 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 anthelminitics 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 Compar46. 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).
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 (synapomorphies; 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 synapomorphies 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 (playethylmintheg tentum or nematode cuticle maintenance proteins) and protein degradation (proteases and protease inhibitors).

Among nematodes, clade IVa (which includes Stronglyloides spp.; Fig. 1) showed the highest number of clade-specific families, including a novel ferrochelatase-like family. Most nematodes lack functional ferrochelatases for the last step of haem biosynthesis, but harbor ferrochelatase-like genes of unknown function, to which the synapomorphic 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 parasitic 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 2.5, Supplementary Table 10 and Methods) revealed intra- and interspecific expansions of SCP/TAPS sequences (Supplementary Note 2.5, Supplementary Fig. 11a,b). A more striking expansions (Supplementary Table 9a and Supplementary Fig. 9d) suggests a role in immune modulation. Repertoires of glycosyl transferases have expanded in nematode clades IV and V, and tapeworms (Supplementary Fig. 9a–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 IVc 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 expressions 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 superfamily 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 superfamily 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.
Fig. 3 | Distribution and phylogeny of SCP/TAPS genes. A maximum-likelihood tree of SCP/TAPS genes. Colors represent different species groups. Homo sapiens GLIPR2 was used to root the tree. Blue dots show high bootstrap values (≥0.8). A clade was collapsed into a triangle if more than half its leaves were genes from the same species group. Nematode clade I had fewer counts, but was collapsed to show its relationship to other clades’ expansions. ‘Strongylid’ refers to clades Va, Vb and Vc.

for existing drugs for human diseases21, and are attractive leads for developing new ones. We therefore explored the diversity of these superfamilies across the nematodes and platyhelminths (Supplementary Fig. 13, Supplementary Note 3 and Methods).

Proteases and protease inhibitors perform diverse functions in parasites, including immunomodulation, host tissue penetration, modification of the host environment (for example, anticoagulation) and digestion of blood75. M12 astacins have particularly expanded in nematode clade IVa (five families), as previously reported18, but there are two additional expansions in clades Vc and Vb (Fig. 4, Supplementary Fig. 14 and 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 tissue76. The cathepsin B C1-cysteine proteases are particularly expanded in species that feed on blood (two expansions in nematode clades Vc and Va19, with highest platyhelminth gene counts in schistosomatids and Fasciola20; Supplementary Fig. 14). Indeed, they are involved in blood digestion in adult nematodes19 and platyhelminths20, but some likely have different roles such as larval development20 and host invasion19.

Different protease inhibitors may modulate activity of parasite proteases or protect parasitic nematodes and platyhelminths from degradation by host proteases, facilitate feeding or manipulate the host response to the parasite17. The I2 (Kunitz-BPTI) trypsin inhibitors are the most abundant protease inhibitors across parasitic nematodes and platyhelminths (Fig. 4). An expansion of the I17 family, which includes secretory leukocyte peptidase inhibitor, was reported previously in Trichuris muris17 but the striking confinement of this expansion to most of the parasites of clade I is now apparent (Fig. 4). We also observed a notable family of α-2-macroglobulin (I39) protease inhibitors that are present in all platyhelminths but expanded in tapeworms (Supplementary Fig. 14). The tapeworm α-2-macroglobulins may be involved in reducing blood clotting at attachment or feeding sites; alternatively, they may modulate the host immune response, since α-2-macroglobulins bind several cytokines and hormones40. Chymotrypsin/elastase inhibitors (family I18) were particularly expanded in clades Vc and IVa (consistent with upregulation of I18 genes in Strongyloides parasitic stages41) and to a lesser extent in clade IIIb (Fig. 4), consistent with evidence that they may protect Ascaris from host proteases41.

We also identified protein domain combinations that were specific to either nematodes or platyhelminths (131 and 50 domain combinations, respectively). Many of these involved protease and protease inhibitor domains. In nematodes, several combinations included Kunitz protease inhibitor domains, and in platyhelminths metalloprotease families M18 and M28 were found in novel combinations (Supplementary Table 14, Supplementary Note 3.2 and Methods).

Of the 230 gene families annotated as GPCRs (Supplementary Figs. 13 and 15 and Supplementary Note 3.3), only 21 were conserved across phyla. Chemosensory GPCRs, while abundant in nematodes, were not identified in platyhelminths, although they are identifiable in other Lophotrochozoa (such as Mollusca82), suggesting that either the platyhelminths have lost this class or they are very divergent (Supplementary Table 15). GPCR families lacking sequence similarity with known receptors included the platyhelminth-specific rhodopsin-like orphan families (PROFs), which are likely to be class A receptors and peptide responsive, and several other fluke-specific non-PROF GPCR families. The massive radiation of chemoreceptors in C. elegans was unmatched in any other nematode (87% versus ≤48% of GPCRs). 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 3.4). For example, glutamate signaling arose independently in platyhelminths and nematodes56, and in trematodes the normal role of acetylcholine has been reversed, from activating to inhibitory57. Our analysis suggested the platyhelminth acetylcholine-gated anion channels are most related to the Acr26/27 group of nematode nicotinic acetylcholine receptors that are the target of the anthelmintics moxidectin and pirantel58, rather than to nematode acetylcholine-gated cation channels, targeted by nicotine and levamisole (Supplementary Fig. 17).

ABC transporters (Supplementary Table 17 and Supplementary Note 3.5) and kinases (Supplementary Note 3.6 and Supplementary Fig. 18) showed losses and independent expansion within nematodes and platyhelminths. The P-glycoprotein class of transporters, responsible for the transport of environmental toxins and linked with anthelmintic resistance, is expanded relative to vertebrates59, with increased numbers in nematodes (Supplementary Fig. 19).

Metabolic reconstructions of nematodes and platyhelminths. In the context of drug discovery, understanding the metabolic capabilities of parasitic worms may reveal vulnerabilities that can be exploited in target-based screens for new compounds. For each of the 81 nematode and platyhelminth species, metabolism was reconstructed based on high confidence assignment of enzyme classes (Supplementary Table 18a). The nematodes had a greater
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 Threonine >0.5 Number of proteases/inhibitors relative to all genes (%) 0 >0 0.04 0.08 0.12 0.16 0.2 >0.25 >0.5 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)\(^{10}\) 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\(^{12}\). The glyoxylate cycle appears to have been lost independently in the filaria and *Trichinella* species (Fig. 5a; M000012), 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\(^{12,27}\) and tapeworms, suggesting they obtain pyrimidines from *Wolbachia* endosymbionts or from their hosts, respectively (Supplementary Table 19f). Similarly, all platyhelminths and some nematodes (especially clade VA and filaria IIIc) appear to lack key enzymes for purine synthesis (Supplementary Table 19g) and rely on salvage instead. However, despite the widespread belief that nematodes cannot synthesize purines\(^{12,27}\), complete or near-complete purine synthesis pathways were found in most members of clades I, IIb and V. Nematodes are known to be unable to synthesize haem, but the pathway was found in platyhelminths, including S. mansoni (despite conflicting biochemical data\(^{12,27}\)) (Supplementary Table 19h and Supplementary Table 20i).

Genomes from the β-oxidation pathway, used to break down lipids as an energy source, were not detected in schistosomes and some cyclophyllidean tapeworms (*Hymenolepis, Echinococcus*: Fig. 5a, M000087; Supplementary Table 19a). These species live in glucose-rich environments and may have evolved to use glucose and glyco- gen as principal energy sources. However, biochemical data suggest they do perform β-oxidation\(^{27}\), 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. Platynematids have high numbers of LDH genes, as do blood-feeding *Ancylostoma* hookworms (Supplementary Fig. 22g). Nematode clades Vc (including *Ancylostoma*) and IIb have expansions of α-glcosidases 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\(^{10}\). The importance of the pathway for clade IIb nematodes was reflected in expanded families encoding two key pathway enzymes PEPCCK 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 (CobB/CbIP) is clade IIb-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 IIb (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{94} (Supplementary Fig. 22f), suggested degradation of propionate, originating from malate dismutation or fermentation in the host’s stomach\textsuperscript{95}. Clade 1 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 alvocidib (flavopiridol), which has anthelmintic activity against *C. elegans*\textsuperscript{96}. 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{97}. 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{98}), will undoubtedly refine the resolution of parasite-specific differences, but our gene family analyses have already revealed expansions and synapomorphies 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.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0262-1.

Received: 17 May 2018; Accepted: 24 September 2018; Published online: 05 November 2018

References

Articles Nature Genetics


64. Wu, C. et al. Mapping the binding between the tetraspanin molecule (Sjc23) of Schistosoma japonicum and human non-immune IgG. PLoS ONE 6, e19112 (2011).


79. Acknowledgements

We thank the Wellcome Trust for financial support. M.B. and M.M. were funded by the Wellcome Trust (095159/Z/09/Z) and the Member States of the European Molecular Biology Organization. T.S. was supported by the Wellcome Trust (095159/Z/09/Z) and the Member States of the European Molecular Biology Organization. T.S. was supported by the Wellcome Trust (095159/Z/09/Z) and the Member States of the European Molecular Biology Organization.

Nature Genetics | www.nature.com/naturegenetics

Author contributions

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0262-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.C., M.M. or M.B.

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

Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018

International Helminth Genomes Consortium


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 Global 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 Virginia, Charlottesville, VA, USA. 16Department of Pathobiological Sciences, University of Wisconsin-Madison, Madison, WI, USA. 17Institute of Immunology and Infection Research, University of Edinburgh, Edinburgh, UK. 18Department of Immunology and Infection Research, University of Edinburgh, Edinburgh, UK. 19Institut de Recherche Agricole pour le Développement, Ngaoandéré, Cameroon. 20Hopkirk Research Institute, AgResearch Ltd, Palmerston North, New Zealand. 21Department of Parasitology, Universidad de la República, Montevideo, Uruguay. 22Department of Southwestern Biology, University of New Mexico, Albuquerque, NM, USA. 23Institute of Infection and Immunity, St. George’s, University of London, London, UK. 24Department of Parasitology, University of Granada, Granada, Spain. 25Museum of Southwestern Biology, University of New Mexico, Albuquerque, NM, USA. 26Institute of Infection and Immunity, St. George’s, University of London, London, UK. 27Department of Parasitology, Universidad Complutense de Madrid, Madrid, Spain. 28Division of Parasitic Diseases and Malaria, Centers for Disease Control and Prevention, Atlanta, GA, USA. 29The Carter Center, Atlanta, GA, USA. 30Wellcome Centre for Molecular Parasitology, University of Glasgow, Glasgow, UK. 31The Carter Center, Atlanta, GA, USA. 32Department of Comparative Biology and Experimental Medicine, University of Edinburgh, Edinburgh, UK. 33Department of Parasitology, Chungbuk National University School of Medicine, Cheongju, Korea. 34Department of Comparative Biology and Experimental Medicine, University of Edinburgh, Edinburgh, UK. 35Biodiversity Research Center, Academia Sinica, Taipei, Taiwan. 36Department of Comparative Biology and Experimental Medicine, University of Edinburgh, Edinburgh, UK. 37Institute of Immunology and Infection Research, University of Edinburgh, Edinburgh, UK. 38Department of Parasitology, Chungbuk National University School of Medicine, Cheongju, Korea. 39Department of Immunology and Infection Research, University of Edinburgh, Edinburgh, UK. 40Institute of Parasitology, Chungbuk National University School of Medicine, Cheongju, Korea. 41Institute of Immunology and Infection Research, University of Edinburgh, Edinburgh, UK. 42BaNG, Institute of Evolutionary Biology and Centre for Immunity, Infection and Evolution, University of Edinburgh, Edinburgh, UK. 43Institute of Parasitology, McGill University, Montreal, Quebec, Canada. 44Department of Biomedical Sciences, Iowa State University, Ames, IA, USA. 45Institute for Global Food Security, Queen’s University Belfast, Belfast, UK. 46Wellcome Centre for Molecular Parasitology, University of Glasgow, Glasgow, UK. 47European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK. 48Program in Molecular Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada. 49Department of Biochemistry and Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. 50Department of Immunology and Infection Research, University of Edinburgh, Edinburgh, UK. 51Institut de Recherche Agricole pour le Développement, Ngaoandéré, Cameroon. 52Hopkirk Research Institute, AgResearch Ltd, Palmerston North, New Zealand. 53Department of Parasitology, Universidad de la República, Montevideo, Uruguay. 54Department of Southwestern Biology, University of New Mexico, Albuquerque, NM, USA. 55Institute of Infection and Immunity, St. George’s, University of London, London, UK. 56Department of Parasitology, University of Granada, Granada, Spain. 57Department of Parasitology, Universidad Complutense de Madrid, Madrid, Spain. 58Division of Parasitic Diseases and Malaria, Centers for Disease Control and Prevention, Atlanta, GA, USA. 59The Carter Center, Atlanta, GA, USA. 60Department of Parasitology, Chungbuk National University School of Medicine, Cheongju, Korea. 61Department of Comparative Biology and Experimental Medicine, University of Edinburgh, Edinburgh, UK.
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 pair-end Illumina libraries were prepared from ~0.1 ng to 5 ng genomic DNA, as described for Strongyloides stercoralis. Where there was insufficient DNA, adapter-ligated material was subjected to ~8 PCR cycles. We used 1–10 ng gDNA or whole genome amplification DNA to generate 3 kb mate-pair libraries, as described for S. stercoralis. If there was insufficient gDNA, whole genome amplification was performed using GenomiPhi v2. Each library was run on a Illumina HiSeq 2000 lane. Short insert pair-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 v.1.3.799. 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 v1.2.03 with SGA-corrected reads. For species with 3 kb mate-pair data, the Velvet assembly was scaffolded using SPSPACE. Contigs were extended, and gaps closed and shortened, using Gapfiller and IMAGE. Short fragment reads were re-mapped to the assembly using SMLAT (see URLs), and unassembled reads assembled using Velvet and this merged with the re-assembled assembly. The assembly was re-assembled using SPSPACE, and consensus base quality improved with icORN. REAP was used to break incorrectly assembled scaffolds/contigs. We carried out manual improvement for Wuchereria bancrofti and D. medinensis.

WSI assembly quality control. Assemblies were screened for contamination using BLAST against vertebrate and invertebrate sequences (see ref. 106). For Antirakis simplex, the assembly contained minor laboratory contamination with S. mansoni, which we removed using BLASTN against S. mansoni.

Assembly completeness. CEGMA v2.4 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 SMLAT (see URLs; indexing –k13 -s4 and mapping –a 0.9 -x 1). For each scaffold of ≥8 kb, median (meds) and mean (med) per-base read-depth were calculated using BEDTools102, and genome-wide depth (med,) calculated as the median med, (ref. 1). For scaffolds, the extra sequence that would be gained by ‘uncollapsing’ repeats was estimated as c, = (med, − med) × l/s, (Supplementary Table 5).

WSI gene prediction. Our pipeline had four steps (Supplementary Fig. 27a). First, repeats were masked. Second, preliminary gene predictions, to use as input for MAKER v2.2.28 were generated using Augustus 2.5.5, SNAP 2013-02-16159, GeneMark-ES 2.3a160, genBlasG161 and RATT162. Third, species-specific ESTs and complementary DNA sequences from INSIDC163, and proteins from related species, were aligned to the genome using BLAST164. Last, EST/protein alignments and gene models were used by MAKER to produce a gene set. McDonnell Genome Institute (MGI) data production. The genomes of six species were sequenced at MGI (Supplementary Tables 1 and 2).

MGI sequencing, assembly and quality control. Genome sequencing was carried out on Illumina and 454 instruments (see ref. 165). The workflow for each assembly is in Supplementary Table 1.

Three kilobase, 8 kb and fragment 454 reads (or Illumina reads) were subject to adaptor removal, quality trimming and length filtering (Supplementary Fig. 26b). Cleaned 454 reads were assembled using Newbler before being scaffolded with an in-house tool Giga, which links contigs based on cDNA evidence. Cleaned Illumina reads were assembled using AllPaths-LG166. The assembly was scaffolded further using an in-house tool Pygap, using Illumina short pair-end sequences; and L_RNA_scaffold167, using 454 cDNA data.

An assisted assembly approach was used for Trichinella nativa, whereby cleaned Illumina 3-bp pair-end sequence data were mapped against the T. spiralis genome using bwa168 (Supplementary Fig. 26b), and the T. nativa residues were substituted at aligned positions (see ref. 169). Adapter sequences and contaminants were identified by comparison to a database of vectors and contaminants, using Megasalt170.

MGI transcriptome sequencing and gene prediction. Transcriptome libraries (Supplementary Table 22) were generated with the Illumina TS stranded protocol, and reads assembled using Trinity (see ref. 171).

Genes were predicted using MAKER172, based on input gene models from SNP173, FGENESH (Soffberry), Augustus174, and aligned messenger RNA, EST, transcriptome and protein data from the same or related species (Supplementary Fig. 27b; see ref. 175).

Blaxter Nematode and Neglected Genomics (BaNG) data production. The genomes of three species were sequenced by BaNG (Supplementary Tables 1 and 2).

Sequencing was performed on Illumina HiSeq 2000 and HiSeq 2500 instruments, using 100 or 125 base, paired-end protocols. Paired-end libraries were generated using the Illumina TruSeq protocol.

Sequence data were filtered and contigs using host reads using blottools176. Cleaned reads were normalized with the khmer software177 using a k-mer of 41, and then assembled with ABYSS (v1.3.3)178, with a minimum of three pairs needed to connect contigs during scaffolding (n = 3) (Supplementary Fig. 26c). Assemblies were assessed using blottools and CEGMA179.

Augustus180 was used to predict gene models, trained using annotations from MAKER181. As hints for MAKER, we used Litomofoidos sigmodontis 454 RNA sequencing data assembled with MIRA182 and Newbler183, and Onchocerca ochengi Illumina RNA sequencing data assembled using Trinity (Supplementary Fig. 27c).

Defining high-quality ‘tier 1’ species. A subset of nematode and platyhelminth genomes, termed ‘tier 1’, was selected that had better-quality assemblies and spanned the major clades (Supplementary Table 4). To choose these, species were selected that (1) had contiguous assemblies (usually N50/ scaffold-count ≥5), and complete proteomes (usually CEGMA partial ≥85%), or (2) that helped to ensure ~30% of the genera in each species group (Analysis group’ in Supplementary Table 4) were represented.

Analysis of repeat content and genome size. For each species, repeat libraries were built using RepeatModeler (see URLs), TransposonPSI (see URLs) and LTRharvest184, and the three libraries merged (see ref. 185). The merged library was used to mask repeats in a species’ genome using RepeatMasker (see URLs; –r). We used SSPACE101 to create a mixed-effect model, the species tree (see Methods) was transformed into a ultrametric tree using PAUP*186, with a small constant added to short branches to ensure no zero-length branches were reconstructed; and outgroup species were removed.

Comparo 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) taxonomy187 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 orthologous present in ≥20 species. To do this, proteins in each ortholog group were aligned using MAFETF6.857188; alignments trimmed using GBLOCKS v0.91b189, concatenated and used to build a maximum likelihood tree using a partitioned analysis in RAxML v7.8.6190, 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 91 species (81 helminths and 10 outgroups) in our Compara database (Methods) and always single-copy. For each family, amino acid sequences were aligned using MAFETF 7.205 (–auto). Each alignment was trimmed using GBLOCKS v0.91b (–d4 = 4 -h3 = 4 -b5 = h), and its likelihood calculated on a maximum-parsimony guide tree for all relatively simple (single-matrix) amino acid substitution models in RAxML v8.0.24191, 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.192). 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 v8.2.8193, with a two-state model and the Lewis method to correct for absence of constant-state observations.

Functional annotation. InterProScan194 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 UniProt195 for orthologs identified from our Compara database (Methods), or based on InterPro196 domains (see ref. 197). Gene ontology (GO) terms were assigned by transferring GO terms from orthologs198, and using InterProScan.
Signal peptides and transmembrane domains were predicted using Phobius<sup>141</sup> v.0.1 and SecretomeP<sup>142</sup> v.1.0. A protein predicted by Phobius to have a transmembrane domain was categorized as ‘membrane-bound’, and non-membrane-bound proteins as ‘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 analyzed using KsFin v0.8.3<sup>143</sup>, by providing InterPro IDs (Methods) and a species tree that had clades III, IV and V as a polytomy (Fig. 2). Synapomorphic families were identified at 25 nodes of interest (Supplementary Table 8), by using Dollo parsimony and requiring a family must contain genes from ≥1 descendant species from each child node of the node of interest, and must not contain other species. Families were filtered to retain those that (1) contained ≥90% of descendant species of the node of interest, and (2) in which >90% of species contained ≥1 gene with a particular InterPro domain.

Candidate lateral gene transfers. Ferrocelataceae families in our Compara database (Methods) were extracted by screening for a Ferrocelataceae (IPR001015) domain. Additional ferrocelataceae were retrieved from NCBI for 17 bacterial taxa (Supplementary Table 8c). Sequences were aligned using MAFFT v7.267 (E-INS-i algorithm)<sup>148</sup> and the alignment trimmed using trimAl<sup>148</sup> (gt 0.006); and ref. 149). To control for fragmented assemblies, we used summed protein length per species (in a family) as a proxy for gene count in these metrics:

1. Coefficient of variation:

\[ c, v., = s / \bar{x} \]

where \( s \) is the standard deviation in summed protein length per species, and \( \bar{x} \) its mean.

2. Maximum Z-score:

\[ Z_{\text{max}} = \max \left( \frac{X_{\text{c}, \text{g}} - \bar{X}}{s_{\text{c}, \text{g}}} \right) \]

where \( T \) is the set of non-overlapping species groups (Analysis group in Supplementary Table 4), \( c \) a group in \( T \), \( i \) refers to a particular species, \( X_{\text{c}, \text{g}} \) the mean of the summed protein length (per species) in \( c \), and \( s_{\text{c}, \text{g}} \) the standard deviation in summed protein length per species in species outside \( c \).

3. Maximum enrichment coefficient:

\[ E_{\text{max}} = \max \left( \frac{X_{\text{c}, \text{g}}}{s_{\text{c}, \text{g}}} \right) \]

To increase reliability, these metrics were calculated by only considering tier 1 species (those with high-quality assemblies; Methods). Our code for calculating metrics is available (see URLs).

SCF/TAPS. SCF/TAPS genes were identified as having Pfam PF00188, or being in a SCF/TAPS family in our Compara database (Methods). Those between 146 aa (shortest C. elegans SCF/TAPS) and 1,000 aa were included in the phylogenetic analysis (Supplementary Table 10). Clusters were detected among sequences from a species group (Analysis group in Supplementary Table 4) using USEARCH<sup>146</sup> (UCCLUST, ai identity cut-off = 0.70), and a consensus sequence generated for each cluster. The consensus sequences were aligned using MAFFT<sup>147</sup> (v7.271, -localpair –maxiterate 2 –retree 1 –bl 45); the alignment trimmed with trimAl<sup>148</sup> (gt 0.006); and a maximum likelihood tree built using FastTreeMP<sup>152</sup> (v2.1.7 SSE3, -wag –gamma).

Proteins historically targeted for drug development. Each nematode/platyhelminth proteome was searched against candidate proteomes using MEROPS batch-BLAST<sup>150</sup> (E < 0.001), and PfamScan<sup>153</sup> was used to identify additional homologues in some species (Supplementary Table 11).

Putative GPCRs, identified from the literature and GO:0004930 annotations in WormBase<sup>154</sup>, were used to identify families in our Compara database (Methods). For each family, HHSuite<sup>155</sup> was used to search Uniprot, SCOPUS, Pfam, and PDB; 200 families hitting >2 databases were deemed actual GPCR families (see ref. 156). Additional families were identified from synapomorphies (Methods) and curation, giving 230 GPCR families (Supplementary Table 15).

To build a phylogenetic tree of ion channels, known genes from C. elegans<sup>157</sup>, Brugia malayi<sup>158</sup>, Haemonchus contortus<sup>159</sup>, Oesophasogomum dentatum<sup>160</sup> and S. mansoni<sup>161</sup> were gathered, and their homologues in Compara families in WormBase ParaSite<sup>162</sup>, Genes with <3 or >8 transmembrane domains (predicted by HHATCH<sup>163</sup>) were discarded. Genes were aligned with MAFFT<sup>148</sup> and the alignment trimmed with trimAl<sup>148</sup>. The phylogeny was inferred with MrBayes3.2<sup>164</sup>. Posterior probabilities were calculated from eight reversible jump Markov chains. Monte Carlo chains over 20,000,000 generations.

Kinase models were taken from Kinomer<sup>165</sup>, and thresholds optimized to detect known C. elegans kinases (see ref. 166). The final thresholds were used to filter HMHER search results (against Kinomer) for nematode and platyhelminth species (Supplementary Table 23).

C. elegans ABC transporter and cys-loop receptor subunits genes were collated from WormBase<sup>141</sup>, to which we added H. contortus acr-26 and acr-27 (absent from C. elegans<sup>141</sup>). Homologs in nematodes and platyhelminths were identified using BLASTP (Supplementary Tables 16 and 17).

GO and InterPro/Pfam annotation enrichment. Counts of proteins annotated with each GO term (or InterPro/Pfam domain) per species were normalized by dividing by the total GO annotations in a particular species. To test for enrichment of a particular GO term in a species group (Analysis group in Supplementary Table 4), we used a Mann-Whitney U test to compare normalized counts in that species group, to those in all other species (Supplementary Table 24).

Metabolism. EC (Enzyme Commission number) predictions for nematodes and platyhelminths were derived by combining DETECT v2.0<sup>167</sup>, PRIAM<sup>168</sup>, KAAS<sup>169</sup> and BRENSDA<sup>170</sup> (see ref. 171). Supplementary Fig. 28 and Supplementary Table 18, and supplemented for the 33 tier 1 species (Methods) by pathway hole-filling using Pathway Tools<sup>172</sup> (v18.5). Comparisons of all 81 species (Supplementary Fig. 20a and Supplementary Table 20) did not include ECs from hole-filling. Lower confidence ECs were inferred using families from our Compara database (Methods). Auxotrophies were predicted using Pathway Tools and BioCyc<sup>173</sup>. To predict carbohydrate-active enzymes, HMMER3 was used to search dbCAN<sup>174</sup> (Supplementary Table 25).

Pathway coverage was the fraction of ECs in a reference pathway that were annotated in a species (see ref. 175). We included pathways for which KEGG had a reference pathway for a nematode/platyhelminth (Supplementary Table 18e). Presence of KEGG modules was predicted using modIFS<sup>176</sup>, and species clustered based on module presence using Ward-linkage, based on Jaccard similarity index<sup>177</sup>. Pathway-specific enzymes were predicted following Taylor et al., <sup>178</sup> using subnetworks of KEGG networks formed by just the enzymes (ECs) we had annotated in each particular species.

Potential anthelmintic drug targets and drugs. Potential drug targets. Nematode and platyhelminth proteins from tier 1 species (with high-quality assemblies; Methods) were searched against single-protein targets from ChEMBL<sup>179</sup> using BLASTP (E ≤ 10<sup>−10</sup>). 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/Oesophagostomum melanoaster homologues had lethal phenotypes (see ref. 175).

Potential new anthelmintic drugs. ChEMBL v21<sup>179</sup> 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. 179).

Our top 15% (249) of highest-scoring worm targets had 292,499 compounds. These were filtered by selecting compounds that (1) co-appeared in a PDB<sup>180</sup> (Protein Data Bank in Europe) structure with the ChEMBL target; or (2) had median pChEMBL > 5; leading 131,452 ‘top drug candidates.’

A ‘diverse screening set’. The 131,452 candidates were placed into 27,944 chemical classes, based on ECFP4 fingerprints (see ref. 181). They were filtered by (1) discarding medicinal chemistry compounds that did not co-appear in a PDB structure with the ChEMBL target, or have median pChEMBL > 7; (2) checking availability for purchase in ZINC<sup>182</sup>; and (3) for each worm target, taking the highest-scoring compound from each class; this gave 5,046 compounds.

Self-organizing map. We constructed a self-organizing map of our diverse screening set plus known anthelmintic compounds (Supplementary Table 21a; see ref. 183), using Kohonen v3.02<sup>184</sup> in R v3.3.0, using a 20 x 20 cell hexagonal, non-territorial


Corresponding author(s): Matthew Berriman
Avril Coghlan
Makedonka Mitreva

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

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).

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

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.
Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- [ ] Life sciences
- [ ] Behavioural & social sciences
- [ ] Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

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

<table>
<thead>
<tr>
<th>Materials &amp; experimental systems</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>[ ] Involved in the study</td>
<td>Involved in the study</td>
</tr>
<tr>
<td>[ ] Unique biological materials</td>
<td></td>
</tr>
<tr>
<td>[ ] Antibodies</td>
<td></td>
</tr>
<tr>
<td>[ ] Eukaryotic cell lines</td>
<td></td>
</tr>
<tr>
<td>[ ] Palaeontology</td>
<td></td>
</tr>
<tr>
<td>[ ] Animals and other organisms</td>
<td></td>
</tr>
<tr>
<td>[ ] Human research participants</td>
<td></td>
</tr>
<tr>
<td>[ ] ChIP-seq</td>
<td></td>
</tr>
<tr>
<td>[ ] Flow cytometry</td>
<td></td>
</tr>
<tr>
<td>[ ] MRI-based neuroimaging</td>
<td></td>
</tr>
</tbody>
</table>

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 or laboratory animals or already dead animals (e.g. from an abattoir). Further details on the samples are given in Supplementary Table 1.