Comparative genomics of the major parasitic worms

International Helminth Genomes Consortium

Abstract

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 compare the genomes of 81 nematode and platyhelminth species, including those of 76 parasites. From 1.4 million genes, we identify gene family births and hundreds of large 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 use a wide-ranging *in silico* screen to identify and prioritise new potential drug targets and compounds for testing. We also uncover lineage-specific differences in core metabolism and in protein families historically targeted for drug development. This is the broadest comparative study to date of the genomes of parasitic and non-parasitic worms. It provides a transformative new resource for the research community to understand and combat the diseases that parasitic worms cause.

Introduction

Over a quarter of the human population is infected with parasitic worms, encompassing many species of nematodes (roundworms) and platyhelminths (flatworms)[1]. Although rarely lethal, infections are typified by long-term chronicity, leading to pain, malnutrition, physical disabilities, retarded development, social stigma associated with deformity, and detrimental effects on family members caring for the afflicted. These diseases of poverty impede economic development and attract little medical research investment. Infections with nematodes and platyhelminths make up eight of the WHO list of 20 most neglected tropical diseases of humans and cost the livestock industry billions of dollars annually[2] through lost production[3]. Nematodes also decimate crop production[4].

A limited spectrum of drugs are available to treat nematode or platyhelminth infections. Resistance is widespread in farm animals[5] and mass drug administration in human populations, using repeated monotherapies, is increasing the risk of resistance to human anthelmintics [6]. Ivermectin, the discovery of which led to a Nobel prize, is used to treat many nematode infections[7] but like other anthelmintics does not prevent re-infection[8]. There are no vaccines for humans, and

very few for animals[9]. The commonly-used plant-parasitic nematicides are environmentally toxic[10], and need urgent replacement.

In nematodes (phylum Nematoda), parasitism has arisen multiple times[11]. Nematoda is part of the Ecdysozoa superphylum that also contains Arthropoda and several smaller phyla[12]. Nematodes have been classified into five major clades (numbered I to V), four of which contain human-infective parasites and are analysed here (Fig. 1). In platyhelminths (phylum Platyhelminthes), nearly all parasitic groups had a single origin[13]. The majority of parasites are cestodes (tapeworms) and trematodes (flukes) in the class Neodermata (Fig. 1). Platyhelminthes is part of the superphylum Lophotrochozoa along with molluscs, annelids and other groups. Comparing the genome biology of parasitism between these two phyla may reveal common strategies employed to subvert host defences and drive disease processes.

In this study, we have combined 36 published genomes with new assemblies for 31 nematode and 14 platyhelminth species into the largest genome comparison to date of parasitic and non-parasitic worms. We have used these data to identify gene families associated with the evolution of 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 (Supplementary Table 1, Extended Data Fig. 1) and predicted 0.8 million new protein-coding genes, with 9,132–17,274 genes per species (5–95% percentile range; Supplementary Table 2, Extended Data Fig. 1, Supplementary Information: Results 1.1, 1.2). We combined these new data with 36 previously published worm genomes, comprising one free-living and ten parasitic platyhelminths, four free-living (including the model *Caenorhabditis elegans*) and 21 parasitic nematodes, and 10 non-parasitic outgroup taxa from other animal phyla (Supplementary Tables 3, 4 and Fig. 1), into a comparative genomics resource of 91 species (Fig. 1).

Genome size varied greatly within each phylum, from 42–700 Mb in nematodes, and 104–1,259 Mb in platyhelminths. Repeat content of the genomes also ranged widely, from 3.8–54.5% (5–95% percentile; Supplementary Table 5). To rank factors driving genome size variation in these species, we built a multiple regression model and identified long terminal repeat transposons, simple repeats, assembly quality, DNA transposable elements, total length of introns, and low complexity sequence as being the most important (Supplementary Information: Results 1.3, Supplementary Table 6). Genome size variation is thus largely due to changes in different non-coding elements, as expected[14], suggesting it is either non-adaptive or responding to selection only at the level of overall genome size.

2

Novel gene families and gene family evolution

We inferred gene families from the predicted proteomes of the 91 species using Ensembl Compara (which identifies families based on all-versus-all BLASTP comparisons[15]). Of the 1.6 million proteins, 1.4 million were placed into 108,351 families, for which phylogenetic trees were built and orthology and paralogy inferred (Supplementary Information: Methods 8, Supplementary Fig. 1, 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 (Extended Data Fig. 2).

Gene families specific to, or with significantly changed membership in, particular parasite clades compared to free-living or host taxa are likely to reflect important changes in parasite biology and may be targeted by new antiparasitic interventions (Supplementary Fig. 1). Parasitic species 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 *C. elegans*. We identified 5881 clade-specific families (synapomorphies) at key nodes in the phylogeny relevant to parasitism (Supplementary Information: Results 2.2, Supplementary Table 8). These were either gene family births, or subfamilies that were so diverged from their homologues that they were identified as separate families. Functional annotation of these synapomorphic 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).

Amongst nematodes, clade IVa (which includes *Strongyloides* spp, Fig. 1) displayed the highest number of clade-specific gene families, including a novel ferrochelatase-like (FeCL) family. Most nematodes lack functional ferrochelatases for the last step of haem biosynthesis[16], but harbour FeCL genes of unknown function, to which the synapomorphic clade IVa family was similar (Supplementary Fig. 2). Exceptions are animal parasites in nematode clades III (e.g. the ascarids and the filaria) and IV that acquired a functional ferrochelatase via horizontal gene transfer[17,18].

Within the parasitic platyhelminths (Neodermata) that includes *Protopolystoma*, trematodes and cestodes, a clade-specific inositol-pentakisphosphate 2-kinase (IP2K) was identified. In some species of *Echinococcus* tapeworms IP2K produces inositol hexakisphosphate nano-deposits in the extracellular wall (the laminated layer) that protects larval metacestodes[19]. The deposits increase the surface area for adsorption of host proteins and may promote interactions with the host[20].

We identified 244 gene family expansions in parasites (Supplementary Table 9a; Supplementary Information: Results 2.3, of which 32 had no functional annotation. Of those with some level of functional information, 25 corresponded to previously described expansions, including 21 associated with biological functions important for parasitism (Fig. 2). A further 43 were placed into major functional classes that historically have been most successful as drug targets (kinases, GPCRs, ion channels and proteases[21]; Supplementary Table 9a).

Among the remaining 144 newly identified expansions were families that interact with host immune systems (Fig. 2). *Taenia* tapeworms and clade V strongylid nematodes (i.e. Va, Vb, and Vc; Fig. 1)

3

contained two expanded families with apyrase domains that may have a role in hydrolysing ATP (a host danger signal) from damaged host tissue[22] (Extended Data Fig. 3a-b). 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[23]. In platyhelminths, we observed expansions of tetraspanin families that are likely components of the host-pathogen interface. Examples have been described that are part of extracellular vesicles released by helminths within their hosts[24]; or that bind the Fc domain of host antibodies[25]; or that are highly immunogenic[26] (Extended Data Fig. 3c-d). In strongylids, especially clade Vc, an expansion of the fatty acid and retinol-binding (FAR) gene family, implicated in host-parasite interaction in plant- and animal-parasitic nematodes[27,28] (Extended Data Fig. 3e), suggests a role in immune modulation. Repertoires of glycosyl transferases have expanded in nematode clades Vc and IV, and tapeworms (Extended Data Fig. 4a-c) and may be used to evade or divert host immunity by modifying parasite surface molecules directly exposed to the immune system[29]; alternatively, the surface glycoprotein may interact with lectin receptors on innate immune cells in an inhibitory manner [30]. 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 (Extended Data Fig. 3f). Similarly, a GH5 glycosyl hydrolase family expanded in schistosomes contained genes with egg-enriched expression[31,32], and may be used by eggs to traverse host tissues such as bladder or intestinal walls (Extended Data Fig. 3g). 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[33], and possibly modulates host lectin-based immune activation (Extended Data Fig. 3h).

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[34], *irg-3*[35]) and clades Va/Vc (lysozyme[36] and the dual oxidase *bli-3*[37]) (Supplementary Fig. 3a-d). In nematode clade IIIb, a family was expanded that contains orthologs of the *Parascaris* coiled-coiled protein PUMA, involved in kinetochore biology[38] (Supplementary Fig. 3e). This expansion may be possibly related to the evolution in this group of chromatin diminution, which results in an increase in the number of chromosomes requiring correct segregation during metaphase[39]. In nematode clade IVa and in *Bursaphelenchus*, an expansion of a steroid kinase family (Supplementary Fig. 3f) is suggestive of novelty in steroid-regulated processes in this group, such as the switch between free-living or parasitic stages in *Strongyloides* [40].

The SCP/TAPS (Sperm-Coating Protein/Tpx/Antigen 5/Pathogenesis-related protein 1) genes are a particularly striking example of a family associated with parasitism through their abundance, secretion and immunomodulatory properties[41] but whose functions remain unclear (Supplementary Information: Results 3.1). There were eight expanded families containing SCP/TAPS but a phylogenetic analysis of the full repertoire of 3,167 sequences revealed intra- and inter-specific expansions and diversification over different evolutionary time-scales (Fig. 3 and Supplementary Fig. 4). In particular, the SCP/TAPS superfamily has expanded independently in nematode clades V (18-381 copies in each species) and in clade IVa parasites (39-166 copies) (Fig. 3, Supplementary Fig. 5, Supplementary Table 10). *Dracunculus medinensis* was unusual in being

4

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 the large adult females.

Half (46.9%) of the gene families lacked functional annotation. These families tended to be smaller than those with annotations (Supplementary Fig. 6) and in many cases, may correspond to families that are so highly diverged that ancestry cannot be traced, but may also reflect the huge breadth of unexplored parasite biology. Family 393312, for example, contained 96 members, 26% of which were predicted to be secreted, and was highly expanded in clade Va nematodes (Supplementary Fig. 7; Supplementary Table 9a). Even when families could be functionally annotated to some extent (e.g. based on a protein domain), discerning their precise biological role was in many cases a challenge. For example, a sulfotransferase gene family that was expanded in flukes compared with tapeworms included the *S. mansoni* locus *Smp_089320* that is implicated in resistance to the drug oxamniquine[42] but the endogenous substrate for this enzyme is unknown (Supplementary Fig. 7j).

Proteins historically targeted for drug development

GPCRs, kinases, proteases and ion channels dominate the list of targets for existing drugs for human diseases[21], and are attractive leads for development of new ones. We therefore explored the diversity of these key superfamilies across the nematodes and platyhelminths (Fig. 4; Supplementary Information: Results 3.2, 3.3, 3.4, 3.6).

Proteases and protease inhibitors perform a number of different functions in parasites, including immunomodulation, host tissue penetration, modification of the host environment (e.g. anticoagulation), and digestion of blood[43]. M12 astacins had particularly expanded in nematode clade IVa (five family expansions), as previously reported[44], but also in clades Vc and Vb (two additional expansions; Fig. 4, Supplementary Fig. 8, Supplementary Table 11). As 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), this was consistent with evidence that astacins are involved in skin penetration and migration through connective tissue[45]. The cathepsin B C1-cysteine proteases had particularly expanded in species that feed on blood (two expansions in nematode clades Vc and Va[46], with highest playthelminth gene counts in schistosomatids and the liver fluke *Fasciola*[47]; Supplementary Fig. 8). Indeed, they are involved in blood digestion in both adult nematodes[48] and playhelminths[49], but some likely have different roles such as larval development[50] and host invasion[51].

Different families of protease inhibitors (PI) may modulate the activity of parasite proteases or protect parasitic nematodes and platyhelminths from degradation by host proteases, facilitate feeding or manipulate the host response to the parasite[52]. The I2 (Kunitz-BPTI) trypsin inhibitors were the most abundant PIs across all parasitic nematodes and platyhelminths (Fig. 4b). An expansion of the I17 family, which includes secretory leukocyte peptidase inhibitor, was reported previously in *Trichuris muris*[53] but the striking confinement of this expansion to most of the parasites of clade I is now apparent (Fig. 4b). We also observed a notable family of α-2-macroglobulin (I39) PIs that were present in all platyhelminths but expanded in tapeworms (Supplementary Fig. 8). The tapeworm α-2-macroglobulins may be involved in reducing blood clotting at attachment or feeding sites; alternatively, they may reduce activity of defensive

5

signalling proteins from the host, since α-2-macroglobulins bind several cytokines and hormones[54]. Chymotrypsin/elastase inhibitors (family I8) were particularly expanded in clades Vc and IVa (consistent with upregulation of I8 genes in *Strongyloides* parasitic stages[44]) and to a lesser extent in clade IIIb (Fig. 4b), consistent with evidence that these inhibitors may protect *Ascaris* from host proteases[55]. In our analyses we also identified numerous protein domain combinations that were specific to either nematodes or platyhelminths (131 and 50 domain combinations, respectively). Many of these protein innovations involved protease and PI domains. In nematodes, several combinations included Kunitz PI domains, and in platyhelminths metalloprotease families M18 and M28 were found in novel combinations (Supplementary Table 14; Supplementary Information: Results 3.7).

Of the 230 gene families annotated as GPCRs (Fig. 4a and Extended Data Fig. 5), 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 Mollusca[56]), 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 species (87% vs ≤48% of GPCRs). All parasitic nematodes possessed chemoreceptors, with the most in clade IVa, including several large families synapomorphic to this clade (Extended Data Fig. 5), perhaps related to their unusual life cycles that alternate between free-living and parasitic forms.

Our analysis of pentameric ligand gated ion channels (pLGICs) suggested that the ancestor of each phylum lacked a complete complement of pLGIC subunit classes (Supplementary Fig. 9, Supplementary Table 16). Divergence between nematode and platyhelminth pLGICs resulted from independent expansion and functional divergence. For example, glutamate signalling arose independently in platyhelminths and nematodes[57], and in trematodes the normal role of acetylcholine has been reversed, from activating to inhibitory[58]. Our analysis suggested the platyhelminth acetylcholine-gated anion channels are most related to the Acr-26/27 group of nematode nicotinic acetylcholine receptors that are the target of the anthelmintics morantel and pyrantel[59], rather than to nematode acetylcholine-gated cation channels, targeted by nicotine and levamisole (Supplementary Fig. 10).

ABC transporters (Supplementary Table 17) showed similar loss and independent expansion between nematodes and platyhelminths. The P-glycoprotein class, responsible for the transport of environmental toxins and linked with anthelmintic resistance, was expanded relative to vertebrates[60], with increased numbers in nematodes (Supplementary Fig. 11).

Metabolic reconstructions of nematode and platyhelminth parasites

We generated metabolic reconstructions for each of the 81 nematode and platyhelminth species, based on high confidence assignment of enzyme classes (Supplementary Table 18a). The nematodes in our dataset clearly had a greater range of annotated enzymes per species than did the platyhelminths (Extended Data Fig. 6a), in part reflecting the paucity of biochemical studies in

platyhelminths. Because variation in assembly quality or divergence from model organisms could bias these inferences, we identified losses of pathways and differences in pathway coverage across different clades (Supplementary Information: Results 4, Fig. 5, Supplementary Fig. 12). Pathways related to almost all metabolic superpathways in KEGG[61] showed significantly lower coverage for platyhelminths (versus nematodes) and filaria (versus other nematodes) (Extended Data Fig. 6b).

In contrast to most animals, nematodes possess the glyoxylate cycle, which enables conversion of lipids to carbohydrates that can be used for biosyntheses (e.g. during early development), and to avert starvation[62]. The glyoxylate cycle appeared 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* had also independently lost alanine-glyoxylate transaminase, which 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 appeared to have been lost independently from filaria and tapeworms, suggesting their GCS is non-functional (Supplementary Table 19e). Additionally, filaria had lost the ability to produce and use ketone bodies, a useful temporary store of acetyl CoA under starvation conditions (Supplementary Table 19b). *Dracunculus medinensis*, the outgroup to the filaria in clade IIIc, does have a free-living L1 stage (Supplementary Table 12) and had not lost the glyoxylate cycle, GCS, or usage of ketone bodies.

Multiple initial steps of pyrimidine synthesis were absent from some nematodes, including all filaria (as previously reported[63]) and tapeworms, suggesting these taxa obtain pyrimidines from their hosts or from *Wolbachia* endosymbionts, in the case of many filaria (Supplementary Table 19f). Platyhelminths and many nematodes (especially clade IVa and filaria) were also devoid of key enzymes in purine synthesis (Supplementary Table 19g) so likely rely on salvage instead. However, in contrast to the widespread belief that nematodes cannot synthesise purines[64,65], 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 synthesise haem[16] but the pathway was present in platyhelminths, including *S. mansoni* (despite conflicting biochemical data[16] (Supplementary Table 19h; Supplementary Table 20i).

Genes from the β-oxidation pathway, used to break down lipids as an energy source, were absent from schistosomes 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[66,67], 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 had high numbers of *LDH* genes, as did blood-feeding *Ancylostoma* hookworms (Supplementary Fig. 13g). Nematode clades Vc (which includes *Ancylostoma*) and IIIb had expansions of α-glucosidases that can break down starch and disaccharides in host food to glucose (Supplementary Fig. 13a). Many nematodes and flatworms use malate dismutation as an alternative pathway for anaerobic ATP production[68]. The importance of the pathway for clade IIIb nematodes was reflected in their gene content. We identified expanded families encoding two key

7

pathway enzymes PEPCK and methylmalonyl CoA epimerase, and the intracellular trafficking chaperone for cobalamin (vitamin B-12), a cofactor for the pathway (Supplementary Fig. 13c-e; Supplementary Table 9a). A second cobalamin-related family (CobQ/CbiP) was clade IIIb-specific and appeared to have been gained from bacteria (Supplementary Fig. 14a). A glutamate dehydrogenase family expanded in clade IIIb (Supplementary Fig. 13h) is consistent with a GABA shunt that helps maintain redox balance during malate dismutation. In clade Va we observed expansion of a gene family belonging to a propionate breakdown pathway [69] (Supplementary Fig. 13f), which may degrade propionate from malate dismutation or fermentation in the host's stomach[70]. Clade I nematodes had an acetate/succinate transporter that appeared to have been gained from bacteria, and is likely to participate in acetate/succinate uptake or efflux (Supplementary Fig. 14b).

Identifying new anthelmintic drug targets and drugs

Many of the 24 WHO-listed anthelmintic drugs are compromised by low efficacy, side-effects or rising resistance in parasite populations. To reveal potential new drug targets and drugs, we developed a pipeline to identify the most promising targets from parasitic nematodes and platyhelminths, and compounds likely to interact with them from the ChEMBL database (Extended Data Fig. 7; Supplementary Information: Results 5). For the top 15% of highest-scoring targets (n=289), including 17 out of 19 known or likely targets for WHO-listed anthelmintics that are represented in ChEMBL, the pipeline identified compounds that are predicted to interact (Supplementary Table 21b). This potential screening set, when collapsed by chemical class, contained 5046 drug-like compounds, including 817 drugs with phase III or IV approval and 4229 medicinal chemistry compounds (Supplementary Table 21d). We used a self-organising map (SOM) to cluster these compounds based on their molecular fingerprints (Fig. 6a). This classification showed that the screening set was significantly more structurally diverse than existing anthelmintic compounds (Fig. 6b).

The 289 targets were further prioritised 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 (Extended Data Fig. 8). 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; Extended Data Fig. 8). These compounds included the phase III drug alvocidib (flavopiridol), which has anthelmintic activity against *C. elegans*[71]. Another example is the target cathepsin B, greatly expanded in nematode clade Va (Supplementary Table 9a), for which we identified several compounds including the phase III drug odanacatib, which has recently been shown to have anthelmintic activity against hookworms[72]. 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, and started from very different ancestral gene sets and physiologies. Despite this, the common selective pressures of adaptation to host gut, blood or tissue environments, the specific nutrient abundances available within hosts, the need to avoid or subvert the host immune system and the acquisition of complex multi-host/environment lifecycles to effect transmission, may have driven adaptations in common biological pathways and therefore common genomic solutions. We have surveyed diverse parasitic worms, with a focus on those infecting humans and livestock. Despite the diversity of the species analysed we identified common themes. Gene family analyses revealed expansions in functional classes that reflect the importance to the parasitic lifestyle. We were able to identify biochemical pathways uniquely absent from, or present in, particular parasite groups. Although independent genomic responses are associated with parasitism, observed commonalities among these may provide a basis for universally-applicable approaches to targeting parasites through drugs and other interventions. From these genomes, we analysed 0.5 million proteins for their promise as new targets for development of sorely needed drugs, and suggest likely compounds that target these.

Supplementary Information

Supplementary Information is available in a separate file.

Acknowledgements

Author contributions

Full author contributions are detailed in a separate 'Supplementary Information' file.

Author information

Sequence data have been deposited in the European Nucleotide Archive (ENA). Assemblies and annotation are available at WormBase and WormBase-ParaSite (http://parasite.wormbase.org/). All have been submitted to GenBank under BioProjects listed in Supplementary Table 20. The authors declare no competing interests. Correspondence to Matthew Berriman (mb4@sanger.ac.uk), Makedonka Mitreva (mmitreva@wustl.edu) and Avril Coghlan (alc@sanger.ac.uk).

Affiliations

Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

Helen Beasley, Hayley M. Bennett, Avril Coghlan, James Cotton, Stephen R. Doyle, Daria Gordon, Bhavana Harsha, Thomas Huckvale, Jane Lomax, Nancy Holroyd, Adam J. Reid, Diogo Ribeiro, Gabriel Rinaldi, Myriam Shafie, Eleanor Stanley, Alan Tracey, Matthew Berriman

McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA

Kymberlie Hallsworth-Pepin, John Martin, Philip Ozersky, Bruce A. Rosa, Rahul Tyagi, Xu Zhang, Makedonka Mitreva

BaNG. Institute of Evolutionary Biology and Centre for Immunity, Infection and Evolution, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT UK

Dominik R. Laetsch, Georgios Koutsovoulos, Sujai Kumar, Gaganjot Kaur, Mark Blaxter

INRA PACA Site de Sophia-Antipolis: Sophia-Antipolis, Provence-Alpes-Cote d'Azur, France

Georgios Koutsovoulos (current address)

The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON, M5G 1X8, Canada

Gaganjot Kaur (current address)

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge, CB10 1SD, UK

Kevin L. Howe, Andrew R. Leach, Prudence Mutowo, Neil Rawlings

Biodiversity Research Center, Academia Sinica, Academia Rd., Nangang Dist., Taipei City 115, Taiwan

Tzu-hao Kuo, Tracy J. Lee, Huei-mien Ke, Isheng Jason Tsai

Department of Life Science, National Taiwan Normal University, Heping E. Rd., Da'an Dist., Taipei City 106, Taiwan

Tracy J. Lee and Isheng Jason Tsai

Biodiversity Program, Taiwan International Graduate Program, Academia Sinica and National Taiwan Normal University, Academia Rd., Nangang Dist., Taipei City 115, Taiwan

Tracy J. Lee and Isheng Jason Tsai

Department of Biomedical Sciences, Iowa State University, Ames, Iowa 50011, USA

Nicolas J. Wheeler, Tim A. Day

Department of Pathobiological Sciences, University of Wisconsin-Madison, 1656 Linden Dr, Madison, WI 53706, USA

Nicolas J. Wheeler, Mostafa Zamanian

Institute of Parasitology, McGill University, Montreal, Quebec H9X 3V9, Canada

Robin N. Beech

11

Department of Biochemistry & Molecular and Medical Genetics, University of Toronto, Program in Molecular Structure and Function, The Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada

John Parkinson, Swapna Lakshmipuram Seshadri

Faculty of Medicine, University of Miyazaki, Miyazaki, 889-1692 Japan

Taisei Kikuchi

Wellcome Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, 120 University Place, Glasgow, G12 8TA, UK

Rick M. Maizels

Department of Parasitology, Faculty of Medicine, University of Indonesia, Jakarta, Indonesia

Felix Partono

Institute of Biodiversity Animal Health and Comparative Medicine, Graham Kerr Building, University of Glasgow, Glasgow G12 8QQ, UK

Simon Babayan

Institute of Immunology and Infection Research, Centre for Immunity, Infection and Evolution, School of Biological Sciences, Edinburgh, UK

Judith E. Allen

NextMove Software Ltd, Innovation Centre, Unit 23, Science Park, Milton Road, Cambridge CB4 0EY, UK

Noel O'Boyle

Chang Gung University, No. 261, Wenhua 1st Road, Guishan District, Taoyuan City, Taiwan 333

Lian-Chen Wang

Institute of Biotechnology, Biochemistry and Molecular Parasitology Group, Department of Parasitology, University of Granada, Campus Univ. Fuentenueva, 18071 Granada, Spain

12

Antonio Osuna, Teresa Cruz-Bustos, Mercedes Gomez Samblas

Departamento de Parasitología, Facultad de Farmacia, Universidad Complutense de Madrid, 28040-Madrid, Spain

Carmen Cuéllar

Institute of Infection and Immunity, St. George's, University of London, SW17 ORE, UK; Facultad de Ciencias Medicas, de la Salud y la Vida, Universidad Internacional del Ecuador, Quito, Ecuador

Philip J. Cooper

Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, G61 1QH, UK

Eileen Devaney

Institute of Immunology and Infection Research, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT UK

Simon Babayan, Yvonne Harcus

Centre for Cardiovascular Science, Queen's Medical Research Institute, 47 Little France Crescent, Edinburgh, EH16 4TJ, UK

Yvonne Harcus (current address)

Veterinary Parasitology, Department of Infection Biology, Institute of infection and Global Health, University of Liverpool, L3 5RF, UK

Jane Hodgkinson

Institut de Recherche Agricole pour le Développement, Regional Centre of Wakwa, BP65 Ngaoundéré, Cameroon

Germanus Bah, Vincent N. Tanya

Division of Parasitic Diseases and Malaria, Centers for Disease Control and Prevention, Atlanta, Georgia; The Carter Center, Atlanta, Georgia, USA

Mark L. Eberhard

13

Department of Physiology, School of Medicine, Showa University, Shinagawa-ku, Tokyo 142-8555, Japan

Kazuhito Asano

Institute of Tropical Diseases and Public Health of the Canary Islands, University of La Laguna, Avda. Fco. Sanchez s/n, 38203, Tenerife, Canary Islands, Spain

Pilar Foronda Rodriguez

Joint Faculty of Veterinary Medicine, Yamaguchi University, 1677-1 Yoshida, Yamaguchi 753-8515, Japan

Hiroshi Sato

Department of Comparative Biology and Experimental Medicine, 3330 Hospital Drive, University of Calgary, Calgary, Alberta, T2N 4N1, Canada

John S. Gilleard

Moredun Research Institute, Edinburgh EH26 0PZ, UK
Jacqueline B. Matthews

Museum of Southwestern Biology, 1 University of New Mexico, MSC03-2020, Albuquerque, New Mexico, 87131, USA

Joseph Cook

Departamento de Parasitología, Facultad de Farmacia, Universidad de Valencia, 46100, Burjassot, Valencia, Spain

Rafael Toledo

Biology Centre CAS, Institute of Parasitology, Branišovská 1160/31, 370 05, Czech Republic

Tomáš Scholz

Institute of Parasitology, University of Zurich, Winterthurerstrasse 266a, 8057 Zurich, CH-Switzerland

Manuela Schnyder

Natural History Museum, Cromwell Rd, Kensington, London SW7 5BD, UK

Fiona Allan, Aidan Emery, Peter D. Olson, David Rollinson

Facultad de Ciencias, Universidad de la República, Iguá 4225, 11400 Montevideo, Uruguay

Estela Castillo

Max Planck Institute for Evolutionary Biology, Dept. Evolutionary Ecology Group, August-Thienemann-Str. 2, 24306 Ploen, Germany

Martin Kalbe

Department of Parasitology, Medical Research Institute and Parasite Resource Bank, Chungbuk National University School of Medicine, Cheongju 362-763, Korea

Keeseon S. Eom

Department of Parasitology, Faculty of Science, Charles University, Viničná 7, 12844 Prague 2, Czech Republic

Petr Horák

Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO 63108, USA

Makedonka Mitreva

Department of Microbiology, Immunology, and Tropical Medicine, School of Medicine and Health Sciences, The George Washington University, Washington DC 20037, USA

John M. Hawdon

Unites States Department of Agriculture, Beltsville Human Nutrition Research Center, Diet, Genomics and Immunology Laboratory, Agricultural Research Service, Beltsville, MD 20705, USA

Joseph F. Urban

United States Department of Agriculture, Agricultural Research Service, Beltsville Agricultural Research Center, Animal Parasitic Diseases Laboratory, Beltsville, MD 20705-2350, USA

Dolores E. Hill, Dante Zarlenga

15

Hopkirk Research Institute, AgResearch Ltd., Palmerston North 4442, New Zealand

Stewart A. Bisset

Institute for Medical Microbiology, Immunology and Parasitology, University of Bonn Medical Center, Sigmund-Freud-Str. 25, Building 320, D-53105 Bonn, Germany

Kenneth Pfarr

Institute of Infection & Global Health, University of Liverpool, Liverpool L3 5RF, UK

Benjamin Makepeace

School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT UK

David W. Taylor

References

1   Disease, G. B. D., Injury, I. & Prevalence, C. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **388**, 1545-1602, doi:10.1016/S0140-6736(16)31678-6 (2016).

2   Morgan ER, C. J., Hendrickx G, Biggeri A, Catalan D, von Samson-Himmelstjerna G, Demeler J, Müller E, van Dijk J, Kenyon F, Skuce P. Global change and helminth infections in grazing ruminants in Europe: impacts, trends and sustainable solutions. *Agriculture* **3**, 484-502, doi:10.3390/agriculture3030484 (2013).

3   Charlier, J., van der Voort, M., Kenyon, F., Skuce, P. & Vercruysse, J. Chasing helminths and their economic impact on farmed ruminants. *Trends Parasitol* **30**, 361-367, doi:10.1016/j.pt.2014.04.009 (2014).

4   Jones, J. T. *et al.* Top 10 plant-parasitic nematodes in molecular plant pathology. *Mol Plant Pathol* **14**, 946-961, doi:10.1111/mpp.12057 (2013).

5   Kaplan, R. M. & Vidyashankar, A. N. An inconvenient truth: global worming and anthelmintic resistance. *Vet Parasitol* **186**, 70-78, doi:10.1016/j.vetpar.2011.11.048 (2012).

6   Furtado, L. F., de Paiva Bello, A. C. & Rabelo, E. M. Benzimidazole resistance in helminths: From problem to diagnosis. *Acta Trop* **162**, 95-102, doi:10.1016/j.actatropica.2016.06.021 (2016).

7   Molyneux, D. H. & Ward, S. A. Reflections on the Nobel Prize for Medicine 2015--The Public Health Legacy and Impact of Avermectin and Artemisinin. *Trends Parasitol* **31**, 605-607, doi:10.1016/j.pt.2015.10.008 (2015).

8   Jia, T. W., Melville, S., Utzinger, J., King, C. H. & Zhou, X. N. Soil-transmitted helminth reinfection after drug treatment: a systematic review and meta-analysis. *PLoS neglected tropical diseases* **6**, e1621, doi:10.1371/journal.pntd.0001621 (2012).

9   Hewitson, J. P. & Maizels, R. M. Vaccination against helminth parasite infections. *Expert Rev Vaccines* **13**, 473-487, doi:10.1586/14760584.2014.893195 (2014).

10   Ntalli, N. G. & Caboni, P. Botanical nematicides: a review. *J Agric Food Chem* **60**, 9929-9940, doi:10.1021/jf303107j (2012).

11   Blaxter, M. L. *et al.* A molecular evolutionary framework for the phylum Nematoda. *Nature* **392**, 71-75, doi:10.1038/32160 (1998).

12   Dunn, C. W., Giribet, G., Edgecombe, G. D. & Hejnol, A. Animal Phylogeny and Its Evolutionary Implications. *Annu Rev Ecol Evol S* **45**, 371-+, doi:10.1146/annurev-ecolsys-120213-091627 (2014).

13   Olson, P. D. & Tkach, V. V. Advances and trends in the molecular systematics of the parasitic Platyhelminthes. *Adv Parasitol* **60**, 165-243, doi:10.1016/S0065-308X(05)60003-6 (2005).

14   Lynch, M., Bobay, L. M., Catania, F., Gout, J. F. & Rho, M. The repatterning of eukaryotic genomes by random genetic drift. *Annu Rev Genomics Hum Genet* **12**, 347-366, doi:10.1146/annurev-genom-082410-101412 (2011).

15   Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**, 327-335, doi:10.1101/gr.073585.107 (2009).

16   Rao, A. U., Carta, L. K., Lesuisse, E. & Hamza, I. Lack of heme synthesis in a free-living eukaryote. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 4270-4275, doi:10.1073/pnas.0500877102 (2005).

17   Wu, B. *et al.* Interdomain lateral gene transfer of an essential ferrochelatase gene in human parasitic nematodes. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 7748-7753, doi:10.1073/pnas.1304049110 (2013).

18   Nagayasu, E. *et al.* Identification of a bacteria-like ferrochelatase in Strongyloides venezuelensis, an animal parasitic nematode. *PloS one* **8**, e58458, doi:10.1371/journal.pone.0058458 (2013).

19   Casaravilla, C. *et al.* Characterization of myo-inositol hexakisphosphate deposits from larval Echinococcus granulosus. *FEBS J* **273**, 3192-3203, doi:10.1111/j.1742-4658.2006.05328.x (2006).

20   Diaz, A., Casaravilla, C., Barrios, A. A. & Ferreira, A. M. Parasite molecules and host responses in cystic echinococcosis. *Parasite Immunol* **38**, 193-205, doi:10.1111/pim.12282 (2016).

21   Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nat Rev Drug Discov* **16**, 19-34, doi:10.1038/nrd.2016.230 (2017).

22   Zarlenga, D. S., Nisbet, A. J., Gasbarre, L. C. & Garrett, W. M. A calcium-activated nucleotidase secreted from Ostertagia ostertagi 4th-stage larvae is a member of the novel salivary apyrases present in blood-feeding arthropods. *Parasitology* **138**, 333-343, doi:10.1017/S0031182010001241 (2011).

23   Cathcart, M. K. & Bhattacharjee, A. Monoamine oxidase A (MAO-A): a signature marker of alternatively activated monocytes/macrophages. *Inflamm Cell Signal* **1**, doi:10.14800/ics.161 (2014).

24   Coakley, G., Maizels, R. M. & Buck, A. H. Exosomes and Other Extracellular Vesicles: The New Communicators in Parasite Infections. *Trends Parasitol* **31**, 477-489, doi:10.1016/j.pt.2015.06.009 (2015).

25   Wu, C. *et al.* Mapping the binding between the tetraspanin molecule (Sjc23) of Schistosoma japonicum and human non-immune IgG. *PloS one* **6**, e19112, doi:10.1371/journal.pone.0019112 (2011).

17

26   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 neglected tropical diseases* **11**, e0005306, doi:10.1371/journal.pntd.0005306 (2017).

27   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).

28   Rey-Burusco, M. F. *et al.* Diversity in the structures and ligand-binding sites of nematode fatty acid and retinol-binding proteins revealed by Na-FAR-1 from Necator americanus. *Biochem J* **471**, 403-414, doi:10.1042/BJ20150068 (2015).

29   Dell, A., Haslam, S. M. & Morris, H. R. in *Parasitic Nematodes: molecular biology, biochemistry and immunology.* (eds M. W. Kennedy & W. Harnett) (Cabi Publishing, 2013).

30   Rodrigues, J. A. *et al.* Parasite Glycobiology: A Bittersweet Symphony. *PLoS Pathog* **11**, e1005169, doi:10.1371/journal.ppat.1005169 (2015).

31   Young, N. D. *et al.* Whole-genome sequence of Schistosoma haematobium. *Nat Genet* **44**, 221-225, doi:10.1038/ng.1065 (2012).

32   Anderson, L. *et al.* Schistosoma mansoni Egg, Adult Male and Female Comparative Gene Expression Analysis and Identification of Novel Genes by RNA-Seq. *PLoS neglected tropical diseases* **9**, e0004334, doi:10.1371/journal.pntd.0004334 (2015).

33   Gong, H. *et al.* A novel PAN/apple domain-containing protein from Toxoplasma gondii: characterization and receptor identification. *PloS one* **7**, e30169, doi:10.1371/journal.pone.0030169 (2012).

34   Parsons, L. M. *et al.* Caenorhabditis elegans bacterial pathogen resistant bus-4 mutants produce altered mucins. *PloS one* **9**, e107250, doi:10.1371/journal.pone.0107250 (2014).

35   Shapira, M. *et al.* A conserved role for a GATA transcription factor in regulating epithelial innate immune responses. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 14086-14091, doi:10.1073/pnas.0603424103 (2006).

36   Hewitson, J. P. *et al.* Proteomic analysis of secretory products from the model gastrointestinal nematode Heligmosomoides polygyrus reveals dominance of venom allergen-like (VAL) proteins. *J Proteomics* **74**, 1573-1594, doi:10.1016/j.jprot.2011.06.002 (2011).

37   van der Hoeven, R., Cruz, M. R., Chavez, V. & Garsin, D. A. Localization of the Dual Oxidase BLI-3 and Characterization of Its NADPH Oxidase Domain during Infection of Caenorhabditis elegans. *PloS one* **10**, e0124091, doi:10.1371/journal.pone.0124091 (2015).

38   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 Parascaris. *J Cell Sci* **111 (Pt 6)**, 723-735 (1998).

39   Tobler, H., Etter, A. & Muller, F. Chromatin diminution in nematode development. *Trends Genet* **8**, 427-432 (1992).

40   Albarqi, 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, doi:10.1371/journal.ppat.1005358 (2016).

41   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, doi:10.1016/j.biotechadv.2009.02.005 (2009).

42   Valentim, C. L. *et al.* Genetic and molecular basis of drug resistance and species-specific drug action in schistosome parasites. *Science* **342**, 1385-1389, doi:10.1126/science.1243106 (2013).

43   McKerrow, J. H., Caffrey, C., Kelly, B., Loke, P. & Sajid, M. Proteases in parasitic diseases. *Annu Rev Pathol* **1**, 497-536, doi:10.1146/annurev.pathol.1.110304.100151 (2006).

44   Hunt, V. L. *et al.* The genomic basis of parasitism in the Strongyloides clade of nematodes. *Nat Genet* **48**, 299-307, doi:10.1038/ng.3495 (2016).

45   Williamson, A. L. *et al.* Ancylostoma caninum MTP-1, an astacin-like metalloprotease secreted by infective hookworm larvae, is involved in tissue migration. *Infect Immun* **74**, 961-967, doi:10.1128/IAI.74.2.961-967.2006 (2006).

46   Laing, R. *et al.* The genome and transcriptome of Haemonchus contortus, a key model parasite for drug and vaccine discovery. *Genome biology* **14**, R88, doi:10.1186/gb-2013-14-8-r88 (2013).

47   McNulty, S. N. *et al.* Genomes of Fasciola hepatica from the Americas Reveal Colonization with Neorickettsia Endobacteria Related to the Agents of Potomac Horse and Human Sennetsu Fevers. *PLoS Genet* **13**, e1006537, doi:10.1371/journal.pgen.1006537 (2017).

48   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, doi:10.1074/jbc.M405842200 (2004).

49   Delcroix, M. *et al.* A multienzyme network functions in intestinal protein digestion by a platyhelminth parasite. *J Biol Chem* **281**, 39316-39329, doi:10.1074/jbc.M607128200 (2006).

50   Duffy, M. S., Cevasco, 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, doi:10.1128/IAI.74.2.1297-1304.2006 (2006).

51   Cancela, M. *et al.* A distinctive repertoire of cathepsins is expressed by juvenile invasive Fasciola hepatica. *Biochimie* **90**, 1461-1475, doi:10.1016/j.biochi.2008.04.020 (2008).

52   Knox, D. P. Proteinase inhibitors and helminth parasite infection. *Parasite Immunol* **29**, 57-71, doi:10.1111/j.1365-3024.2006.00913.x (2007).

53   Foth, B. J. *et al.* Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction. *Nat Genet* **46**, 693-700, doi:10.1038/ng.3010 (2014).

54   Rehman, A. A., Ahsan, H. & Khan, F. H. alpha-2-Macroglobulin: a physiological guardian. *J Cell Physiol* **228**, 1665-1675, doi:10.1002/jcp.24266 (2013).

55   Martzen, M. R., Geise, G. L., Hogan, B. J. & Peanasky, R. J. Ascaris suum: localization by immunochemical and fluorescent probes of host proteases and parasite proteinase inhibitors in cross-sections. *Exp Parasitol* **60**, 139-149 (1985).

56   Nei, M., Niimura, Y. & Nozawa, M. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. *Nat Rev Genet* **9**, 951-963, doi:10.1038/nrg2480 (2008).

57   Lynagh, T. *et al.* Molecular basis for convergent evolution of glutamate recognition by pentameric ligand-gated ion channels. *Sci Rep* **5**, 8558, doi:10.1038/srep08558 (2015).

58   MacDonald, K. *et al.* Functional characterization of a novel family of acetylcholine-gated chloride channels in Schistosoma mansoni. *PLoS Pathog* **10**, e1004181, doi:10.1371/journal.ppat.1004181 (2014).

59   Courtot, E. *et al.* Functional Characterization of a Novel Class of Morantel-Sensitive Acetylcholine Receptors in Nematodes. *PLoS Pathog* **11**, e1005267, doi:10.1371/journal.ppat.1005267 (2015).

60   Vasiliou, V., Vasiliou, K. & Nebert, D. W. Human ATP-binding cassette (ABC) transporter family. *Hum Genomics* **3**, 281-290 (2009).

61   Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* **40**, D109-114, doi:10.1093/nar/gkr988 (2012).

19

62   Kondrashov, F. A., Koonin, E. V., Morgunov, I. G., Finogenova, T. V. & Kondrashova, M. N. Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation. *Biol Direct* **1**, 31, doi:10.1186/1745-6150-1-31 (2006).

63   Godel, C. *et al.* The genome of the heartworm, Dirofilaria immitis, reveals drug and vaccine targets. *FASEB J* **26**, 4650-4661, doi:10.1096/fj.12-205096 (2012).

64   Harder, A. The Biochemistry of Haemonchus contortus and Other Parasitic Nematodes. *Adv Parasitol* **93**, 69-94, doi:10.1016/bs.apar.2016.02.010 (2016).

65   Marr, J. J. & Müller, M. *Biochemistry and Molecular Biology of Parasites*. 90 (Academic Press, 1995).

66   Pearce, E. J. & Huang, S. C. The metabolic control of schistosome egg production. *Cell Microbiol* **17**, 796-801, doi:10.1111/cmi.12444 (2015).

67   Ward, C. W. & Fairbairn, D. Enzymes of beta-oxidation and the tricarboxylic acid cycle in adult hymenolepis diminuta (Cestoda) and Ascaris Lumbricoides (Nematoda). *J Parasitol* **56**, 1009-1012 (1970).

68   Mehlhorn, H.   (Springer, 2008).

69   Watson, E. *et al.* Metabolic network rewiring of propionate flux compensates vitamin B12 deficiency in C. elegans. *Elife* **5**, doi:10.7554/eLife.17670 (2016).

70   Van Soest, P. J. *Nutritional ecology of the ruminant.*, (Cornell University Press, 1994).

71   Taylor, C. M. *et al.* Using existing drugs as leads for broad spectrum anthelmintics targeting protein kinases. *PLoS Pathog* **9**, e1003149, doi:10.1371/journal.ppat.1003149 (2013).

72   Vermeire, J. J., Suzuki, B. M. & Caffrey, C. R. Odanacatib, a Cathepsin K Cysteine Protease Inhibitor, Kills Hookworm In Vivo. *Pharmaceuticals (Basel)* **9**, doi:10.3390/ph9030039 (2016).

20

Figures
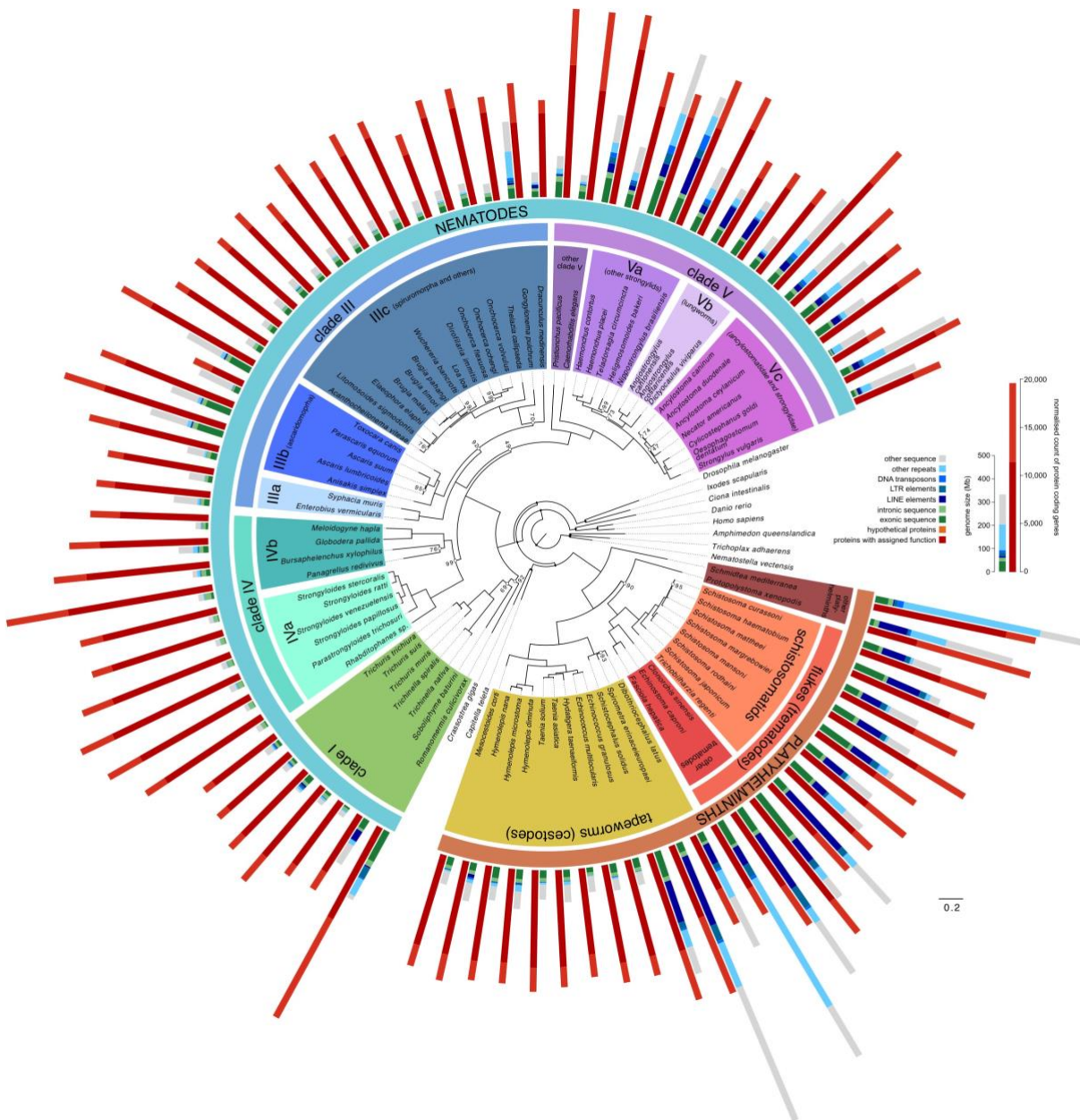
Fig. 1: Genome-wide phylogeny of 81 nematode and platyhelminth species and 10 outgroup species. Maximum-likelihood phylogeny based on a partitioned analysis of a concatenated data matrix of 21,649 amino acid sites from 202 single-copy orthologous proteins present in at least 23 of the species. Values 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 normalised gene count (Supplementary Information: Results 1.2) for proteins with inferred functions based on sequence similarity (having an assigned protein name; Supplementary Information: Methods 3), or those without (named 'hypothetical protein').

21
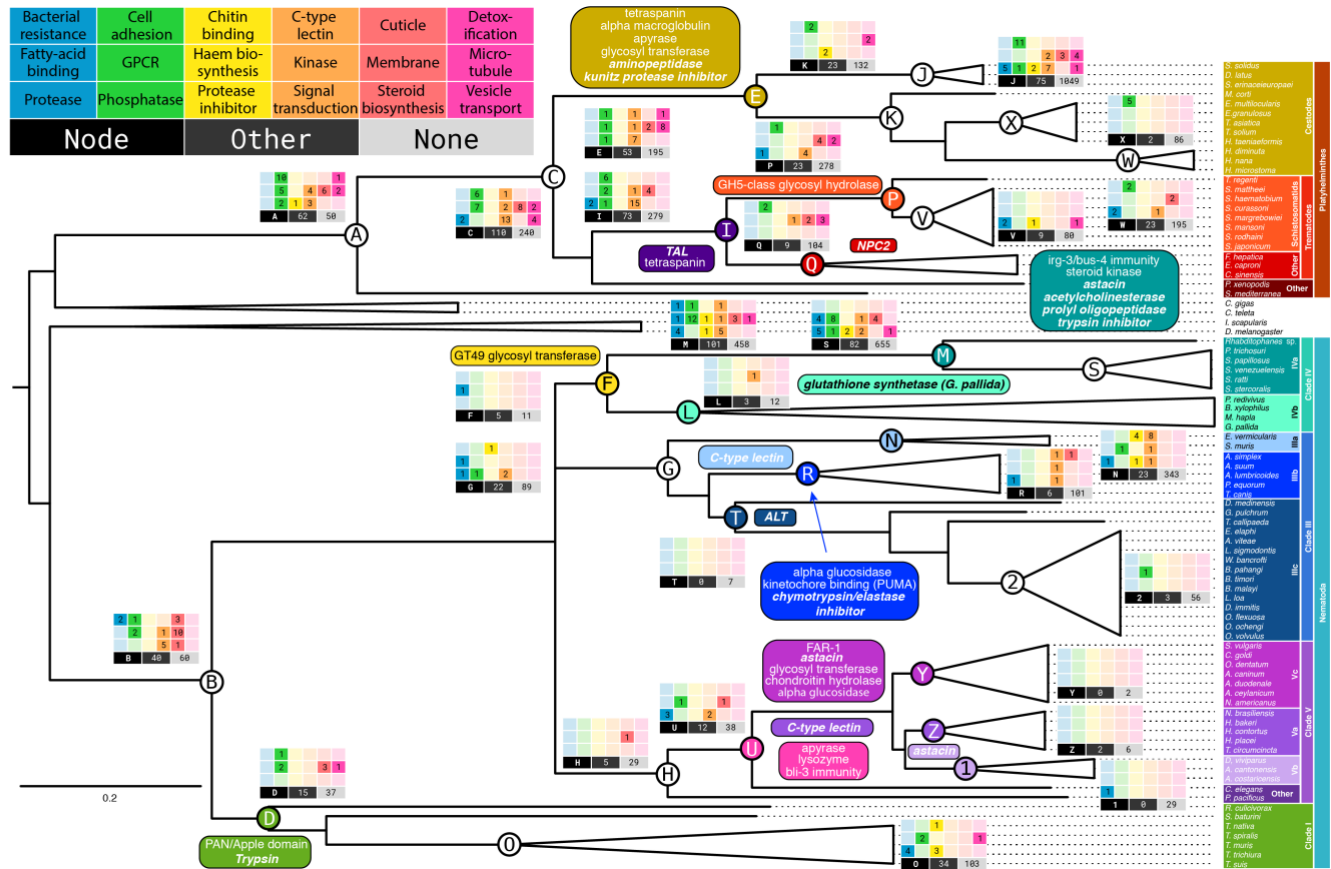
Fig. 2: Functional annotation of synapomorphic and expanded gene families. Rectangular matrices indicate counts of synapomorphic families grouped by 18 functional categories, detailed in the major panel. Representative functional annotation of a family was inferred if more than 90% of the species present 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. Curved panels indicate gene families showing noteworthy expansions at nodes. Previously published expansions are in bold Italics.

1
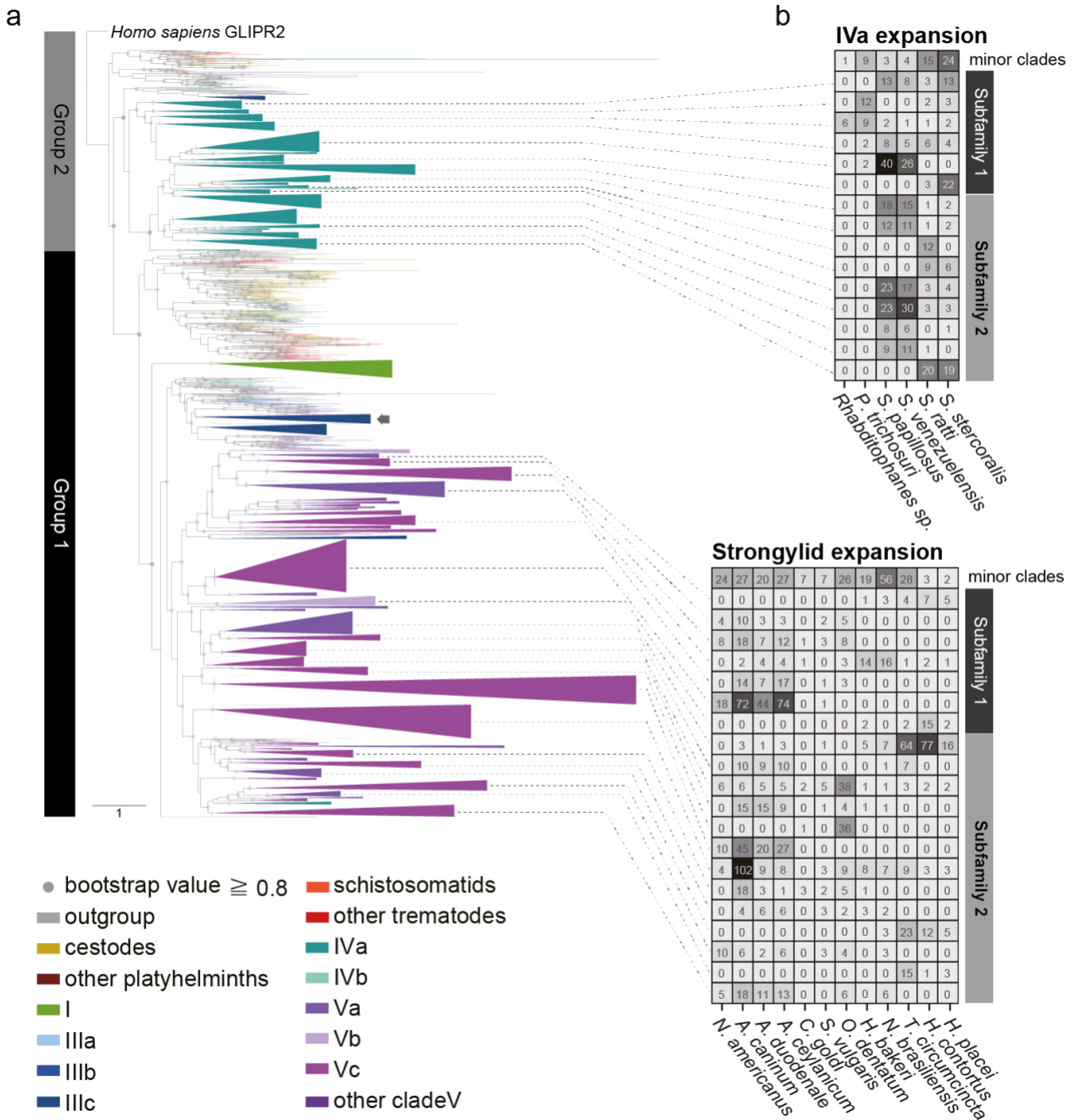
Fig. 3: Distribution and phylogeny of SCP/TAPS genes.

(a) A maximum-likelihood tree of SCP/TAPS genes. Colours represent different species groups. *H. sapiens* GLIPR2 was used to root the tree. Grey 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. Clade I had fewer counts, but was collapsed to show its relationship to other clades' expansions. A *Dracunculus* expansion is indicated with an arrow. (b) Clades containing more than 15 (in group 1) and 10 (in group 2) genes are shown in a heatmap. 'Strongylid' refers to clades Va, Vb and Vc.

1

Fig. 4: Abundances of superfamilies historically targeted for drug development.

(a) Abundances (relative to the total numbers of predicted genes) of genes belonging to superfamilies of interest. Comparisons are (i) platyhelminths vs nematodes, (ii) trematodes vs cestodes and (iii) nematode clade vs other nematodes. * $P<0.05$, ** $P<0.01$, *** $P<10^{-3}$, **** $P<10^{-4}$, ***** $P<10^{-5}$ (Mann Whitney U test). The *C. elegans* GPCR value of 4.6% was not shown. (b) Relative abundance profiles for 84 protease and 31 protease inhibitor families represented in at least 3 of the 81 nematode and platyhelminth species. 33 protease families and 6 protease inhibitor families present in fewer than 3 species were omitted from the visualisation.

3
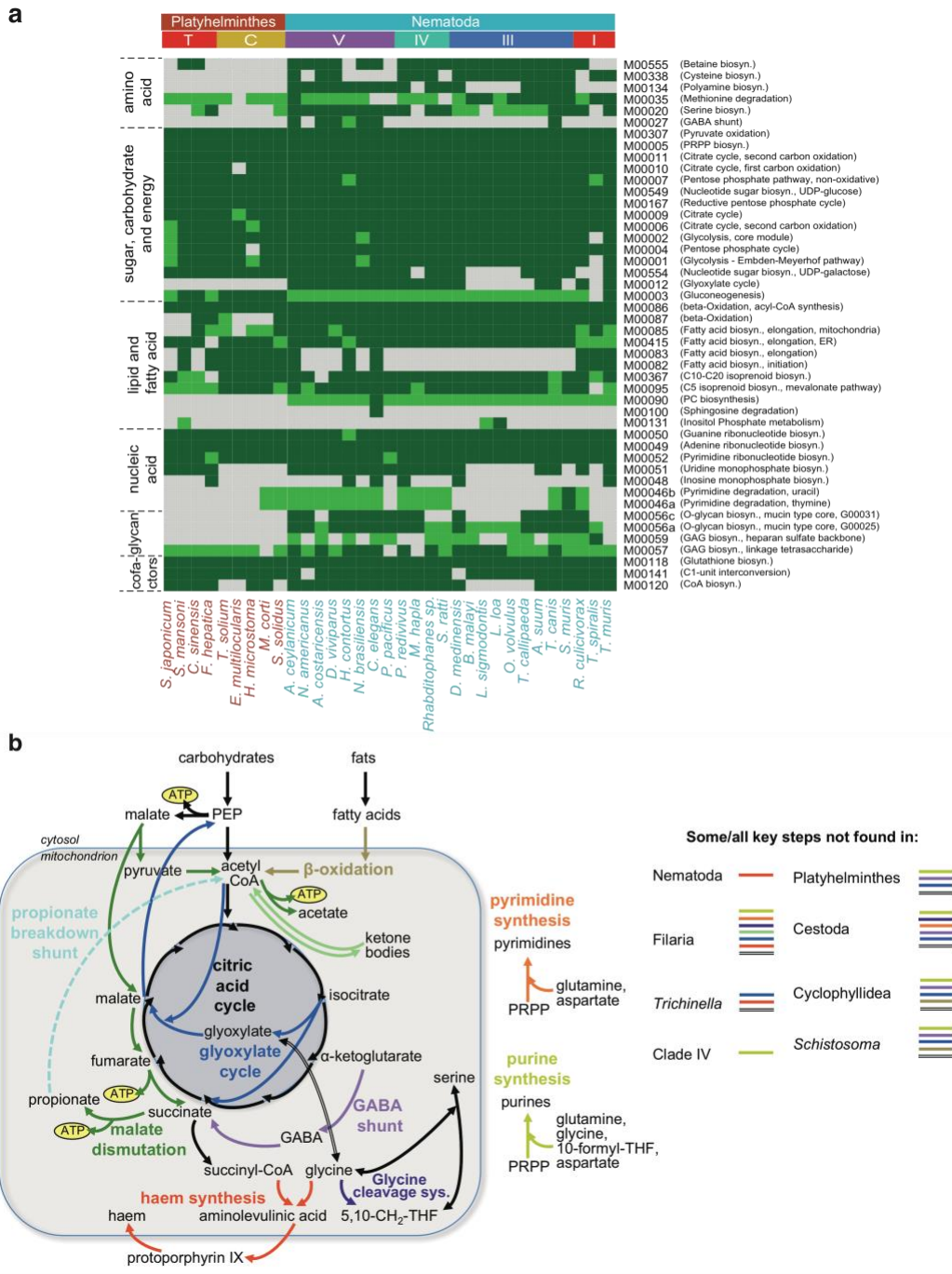
Fig. 5: Metabolic modules and biochemical pathways in platyhelminths and nematodes.
(a) Topology based detection of KEGG metabolic modules among tier 1 (with high quality assemblies; see Supplementary Information: Methods 7) species (dark green = present; light green = largely present (only one enzyme not found). Only modules detected to be complete in at least one species are shown. The EC annotations used for this figure included those from pathway hole-filling and those based on Compara families (Supplementary Table 18a, b). (b) Biochemical pathways that appear to have been completely or partially lost from certain platyhelminth and nematode clades.

4
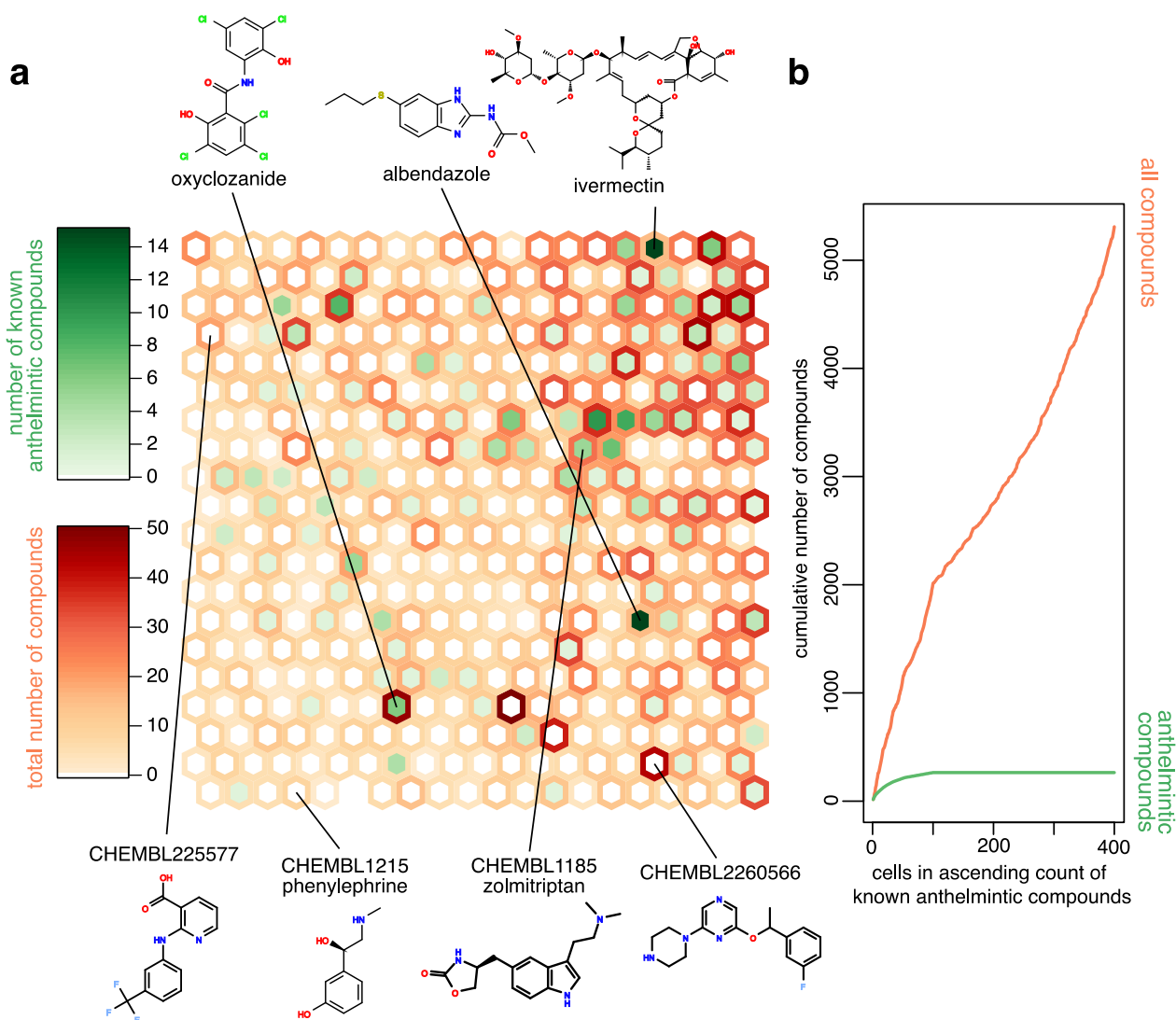
Fig. 6: Self-organising map of known anthelmintic compounds and our screening set of 5046 drug-like compounds.

(a) A self-organising map (SOM) clustering known anthelmintic compounds (Supplementary Table 21a) and our screening set of 5046 compounds. The density of red and green shows the number of screening set and known anthelmintic compounds clustered in each cell, respectively. Structures for representative known anthelmintic compounds are shown at the top, and for screening set compounds along the bottom. (b) Cumulative counts of known anthelmintic and screening set compounds, showing that known anthelmintic compounds are more highly clustered in the SOM. Thus, our screening set represents a more diverse sample of chemical space than the known anthelmintic compounds.

5
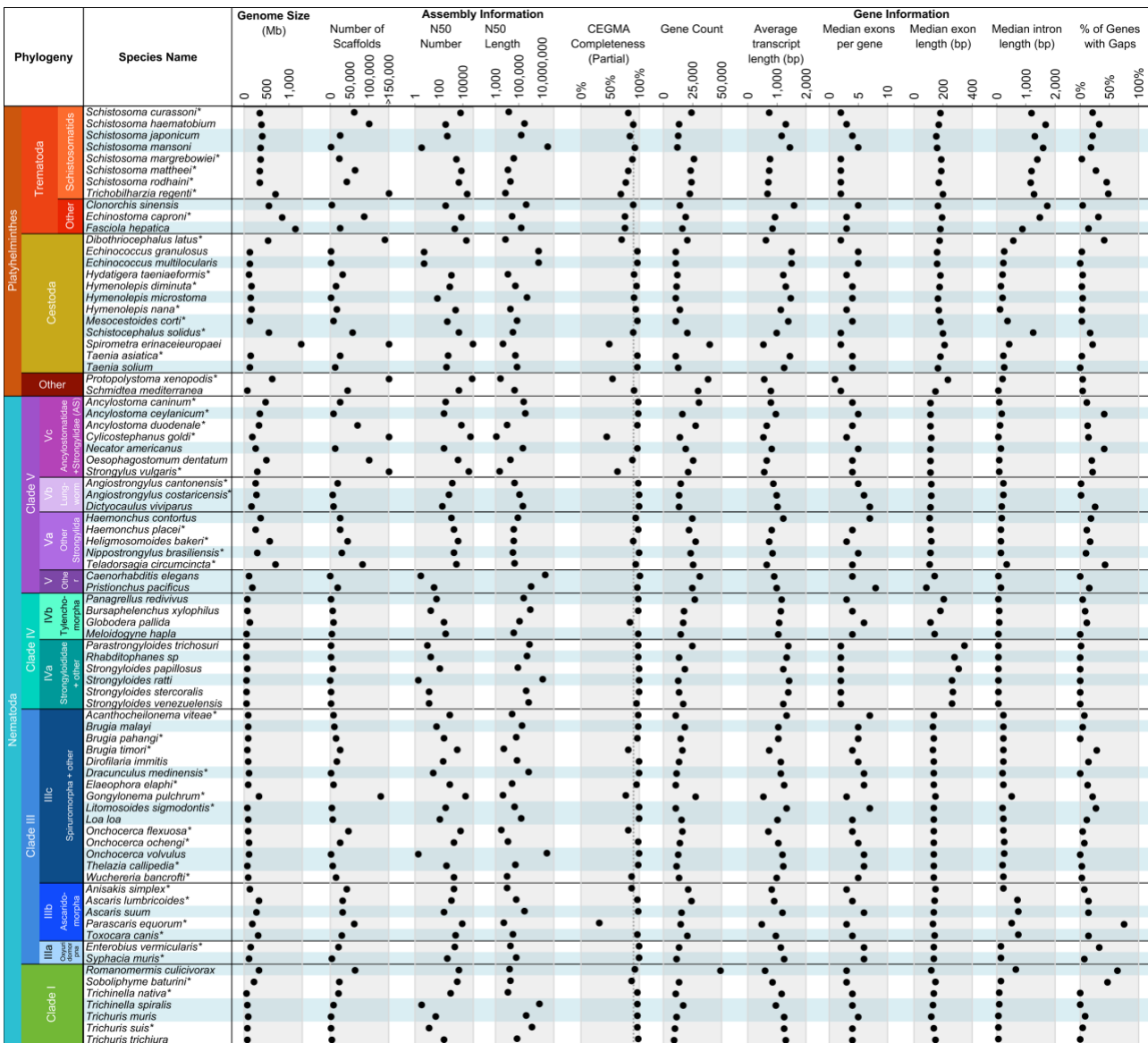
Extended Data Fig. 1: Assembly and gene set statistics.

Blue rows indicate the 33 'tier 1' genomes (with high quality assemblies; Supplementary Information: Methods 7). Asterisks indicate the species for which we have sequenced genomes. The '% of genes with gaps' indicates the proportion of predicted genes which contain a sequence gap (i.e., the gene spans multiple contigs).
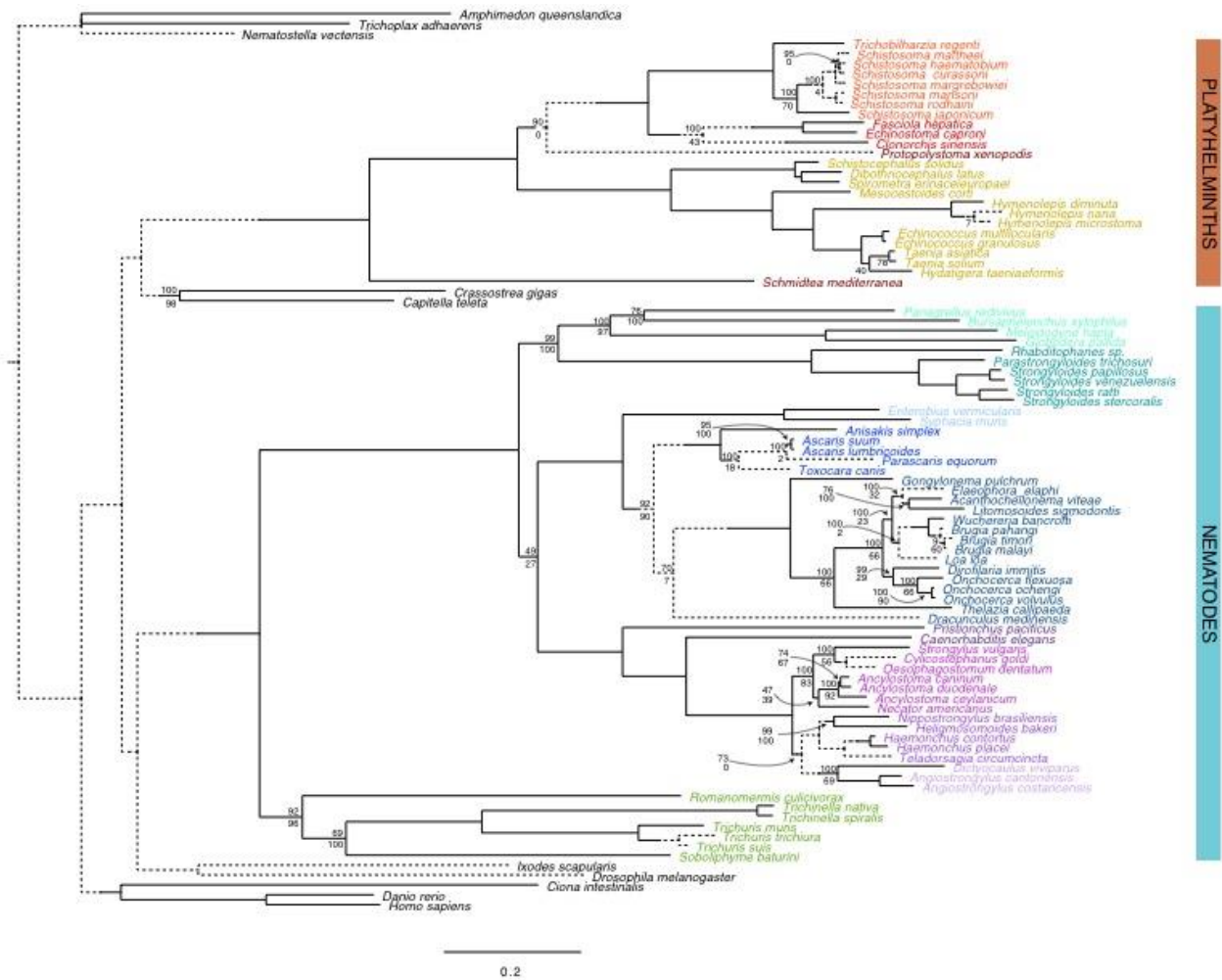
Extended Data Fig. 2: Support for species tree relationships by gene family content.
The phylogenetic tree shown is the species tree from single-copy families, also shown in Fig. 1. Here, branches around nodes not supported by gene family content data are shown with broken lines. Values above nodes are bootstrap values from the amino acid matrix analysis; those below nodes are bootstrap values from the gene content data analysis. Where values are not shown next to solid branches, both are 100/100. Where values are not shown next to dotted branches, they are 100 for the amino acid data and zero for the gene content analysis. Contentious issues include the relationships between the monogeneans, cestodes and trematodes within Platyhelminthes; between nematode clades III, IV and V; between Spiruromorpha+other, Ascaridomorpha and Oxyuridomorpha within clade III; and the positions of the genera *Bursaphelenchus* and *Dictyocaulus* within Nematoda (Supplementary Information: Results 2.6).

7
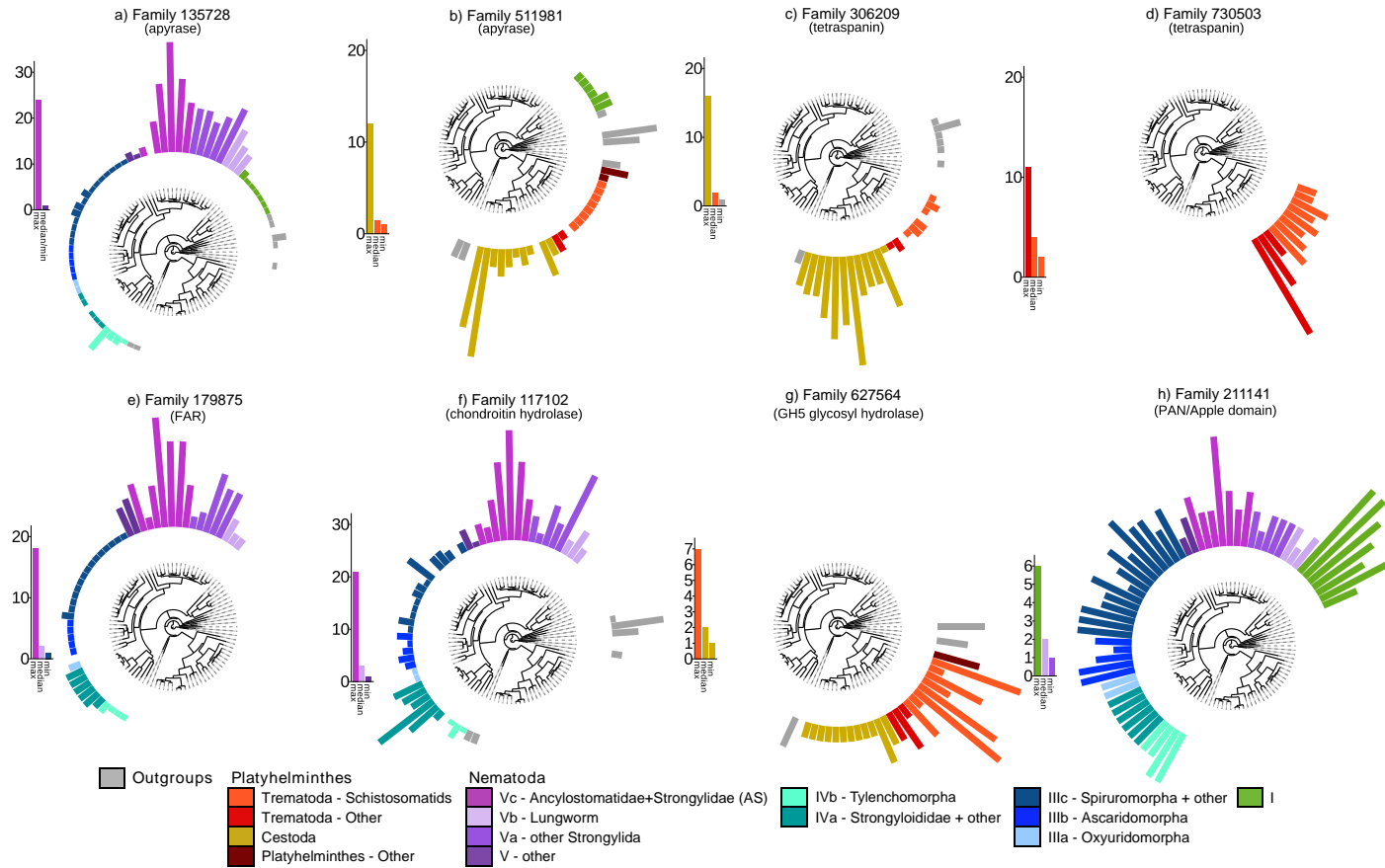
Extended Data Fig. 3: Expanded families involved in parasitism.
Gene families with striking variation across species and potential roles in parasitism. Families were defined using Compara. For colour key and species labels, see Fig. 1. The plot for a family shows the gene count in each species, superimposed upon 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.

1
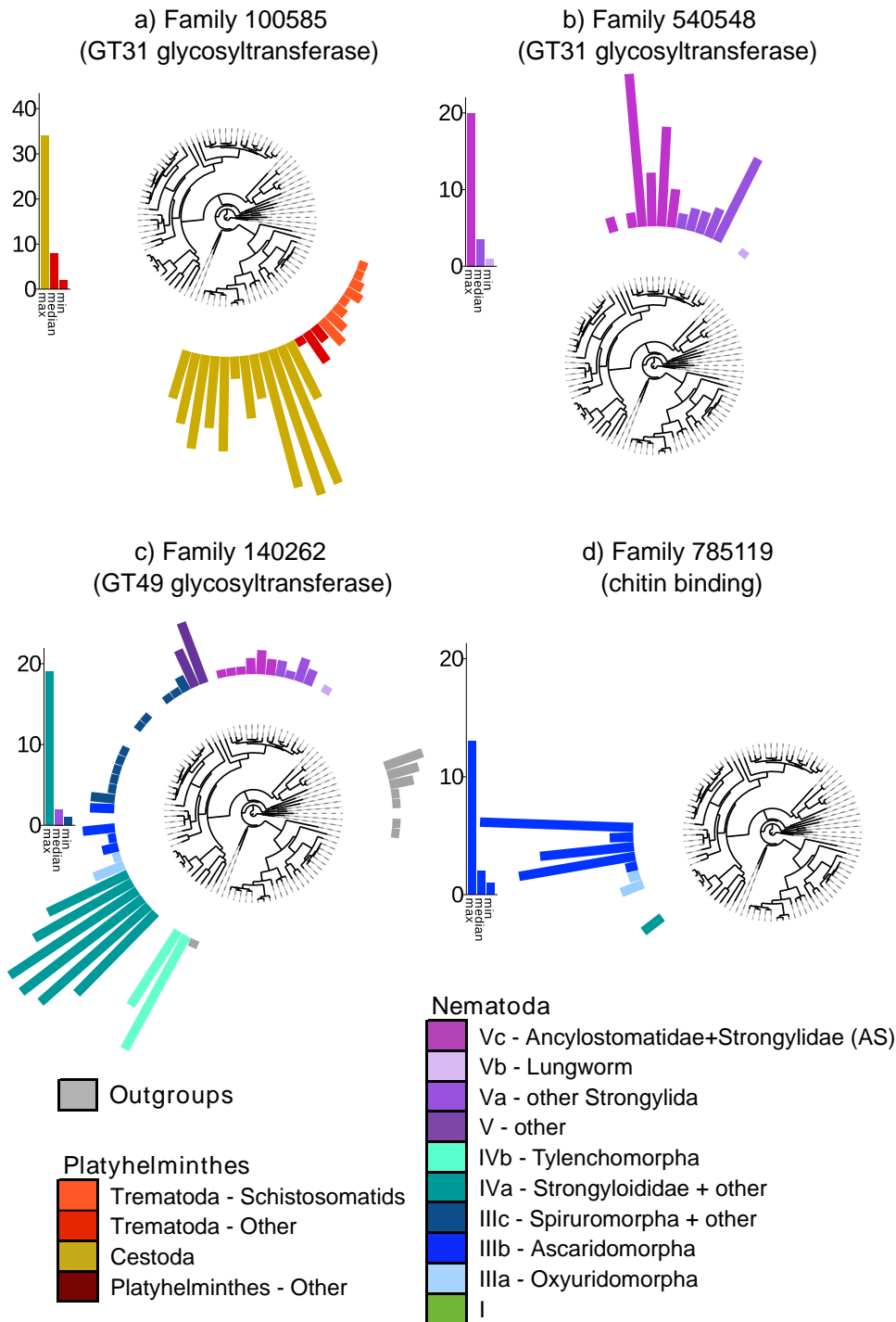
Extended Data Fig 4: Surface coat expanded families.

Expanded families that may be involved in the surface coat. The plot for a family shows the gene count in each species, superimposed upon our species tree. A scale bar beside the plot for each family shows the minimum, median, and maximum gene count across the species, for that family.
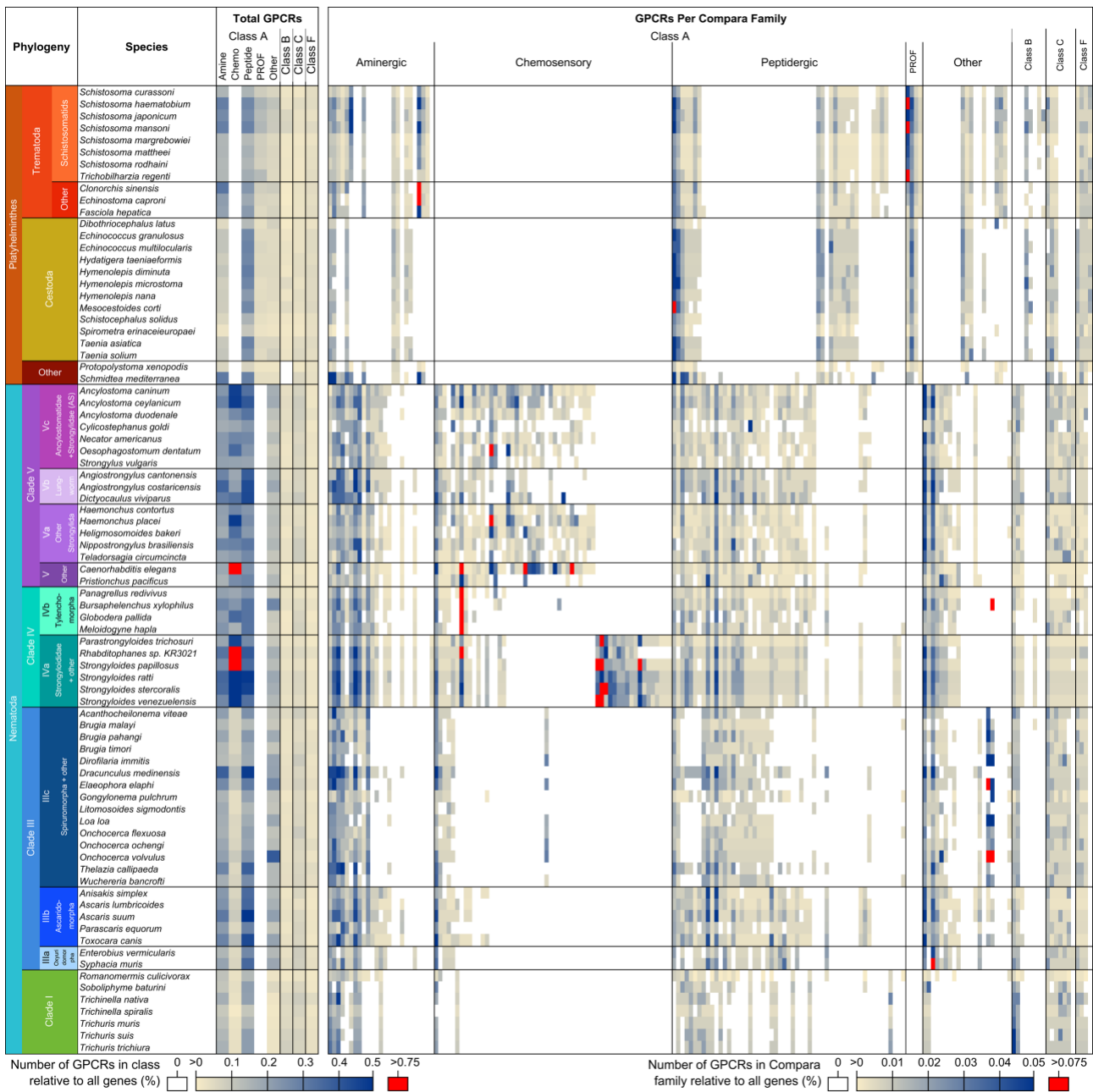
Extended Data Fig. 5: Abundances of GPCRs.

Relative abundance profiles for the major categories of GPCRs and the 180 GPCR gene families (columns) represented in at least three platyhelminth or nematode species (Supplementary Information: Methods 19). 50 families present in fewer than three species were omitted from the visualisation.
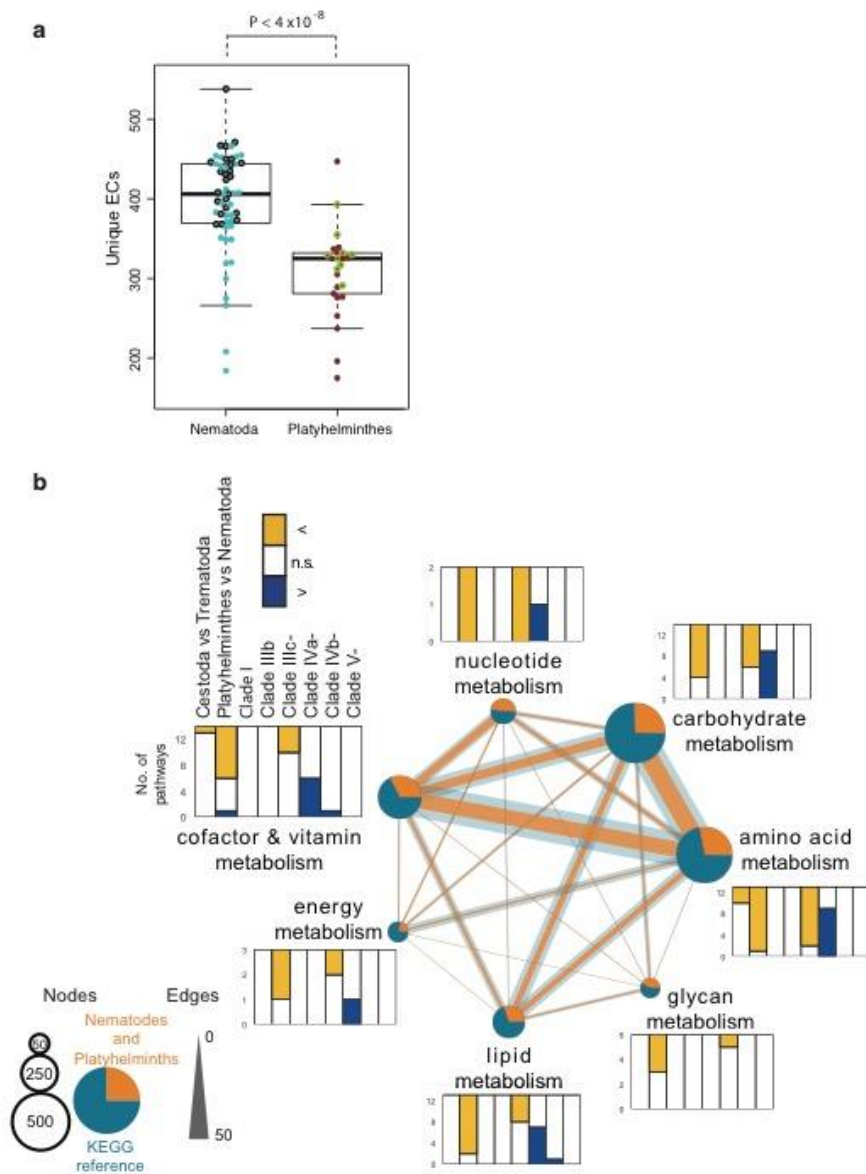
Extended Data Fig. 6: Comparison of EC number annotations and KEGG pathway coverage.
(a) Unique enzymes (EC numbers) annotated before using pathway hole-filling. Tier 1 species (with high quality assemblies; Supplementary Information: Methods 7) are highlighted with thick borders. (b) Comparison of KEGG reference pathway coverages (Supplementary Information: Methods 25), using Wilcoxon tests. The pathways are arranged according to superpathways. Nematode parasite groups are compared with the other five nematode sets combined. The total number and overlap in EC numbers of different superpathways are depicted in the central network. The colour of pies and lines indicate the ECs detected in platyhelminths and nematodes, versus the rest of the KEGG reference database. The size of pie charts and lines indicate the amount of EC numbers. The nematode set names are defined as follows: clade I, clade I nematodes; IIIb, clade IIIb nematodes; IIIc-, filaria (=clade IIIb minus 3 non-filaria); IVa-, parasites of clade IVa (i.e. *Rhabditophanes* excluded); IVb-, parasites of clade IVb (i.e. *P. redivivus* excluded); clade V-, parasites of clade V (i.e. *C. elegans* and *P. pacificus* excluded).

3

Identification of potential anthelmintic drug targets and drugs

Ran BLAST between ~528,000 proteins from 33 'tier 1' helminths and ~6200 single-protein targets in ChEMBL21

Identified ~106,000 helminth genes with top BLAST hits ($E \leq 10^{-10}$) to ~4000 single-protein ChEMBL. Collapsed helminth genes in the same gene family, reducing the target list to **1925 helminth proteins**.

Criteria to create a **target score** (Range, Weight):

- Strength of BLAST match (0—1, 50)
- Lack of human homologs (0—1, 40)
- Predicted lethal or sterile phenotypes (0—1, 30)
- Predicted to be a chokepoint enzyme (0—1, 15)
- Expression in key life cycle stages (0—1, 10)
- Found in multiple helminth species (0—1, 10)
- Has homologs across a major helminth clade (0—1, 10)
- ChEMBL target from a non-chordate animal (0—1, 10)
- Lack of within-species paralogs (0—1, 5)
- Has a highly conserved protein sequence (0—1, 5)
- Matches a ChEMBL target with PDB structure (0—1, 5)

Target scores

827,889 ChEMBL molecules with activity for 1925 targets

Considered the **top 15% (289) of targets**.
449,627 pairs (**286 targets, 292,499 compounds**).

Filtered target-compound pairs using two filters:
(i) the compound & target co-appear in the PDB,
(ii) they have high potency/affinity (high pCHEMBL score) pChEMBL >5 for phase III/IV, >7 for other compounds.
Leaves **52,154 'top drug candidates'** : 595 phase III/IV drugs (in 730 chemical classes), 51,195 others (13,191 classes)

Filtered by **availability for purchase**, using ZINC
Left 6432 compounds in 4543 chemical classes

Scored compounds to take the highest-scoring **from each chemical class, for each target**:

- Drug-likeness (QED score) (0—1, 5)
- Phase of drug approval (phases 0-IV) (0—4, 5)
- Can be administered orally or topically (0—1, 5)
- Has a PDB structure (0—1, 5)
- Penalty for serious side effects (-1—0, 5)
- Penalty for predicted toxicology targets (-1—0, 2.5)

'Diverse screening set':
817 phase III/IV drugs in 661 chemical classes
4229 other compounds in 3936 chemical classes

Identified 40 **'high priority targets'**

'Smaller diverse screening set':
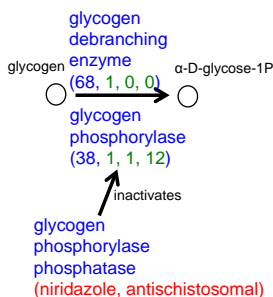**181 phase III/IV drugs** (in 173 chemical classes) **& 539** other compounds (in 530 classes)

Extended Data Fig. 7: Flowchart of the methods used for identifying potential new anthelmintic targets and drugs using ChEMBL.
The following images were used from Openclipart: 'prescription-bottle-and-pills' (by algotruneman), 'famous-and-infamous-molecules-13' (by Firkin) and 'worm-gusano' (by ainara14).
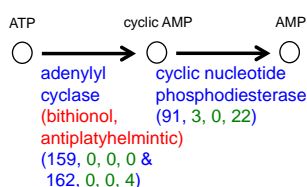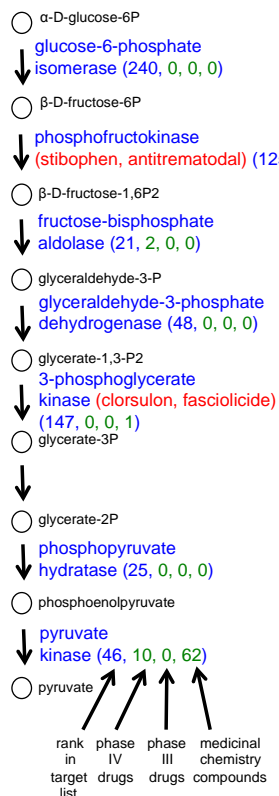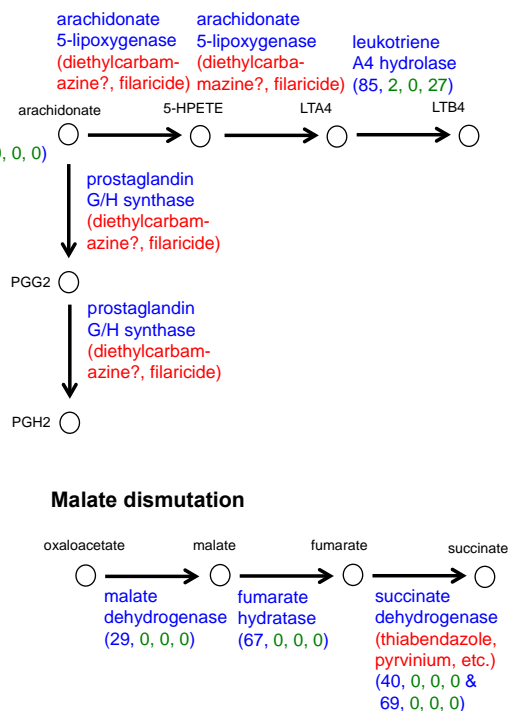
4

Glutathione metabolism

glycine, gamma-L-glutamyl-L-cysteine

glutathione synthetase (4, 0, 0, 0)

glutathione disulphide — glutathione — R-S-glutathione

glutathione reductase (thiacetarsamid, filaricide) (87, 0, 0, 1 & 268, 0, 0, 0)

glutathione transferase (3, 0, 1, 9)

Glycogen metabolism

glycogen debranching enzyme (68, 1, 0, 0)

glycogen — α-D-glycose-1P

glycogen phosphorylase (38, 1, 1, 12)

inactivates

glycogen phosphorylase phosphatase (niridazole, antischistosomal)

Purine metabolism

ATP — cyclic AMP — AMP

adenylyl cyclase (bithionol, antiplatyhelmintic) (159, 0, 0, 0 & 162, 0, 0, 4)

cyclic nucleotide phosphodiesterase (91, 3, 0, 22)

Glycolysis

α-D-glucose-6P

glucose-6-phosphate isomerase (240, 0, 0, 0)

β-D-fructose-6P

phosphofructokinase (stibophen, antitrematodal) (125, 0, 0, 0)

β-D-fructose-1,6P2

fructose-bisphosphate aldolase (21, 2, 0, 0)

glyceraldehyde-3-P

glyceraldehyde-3-phosphate dehydrogenase (48, 0, 0, 0)

glycerate-1,3-P2

3-phosphoglycerate kinase (clorsulon, fasciolicide) (147, 0, 0, 1)

glycerate-3P

glycerate-2P

phosphopyruvate hydratase (25, 0, 0, 0)

phosphoenolpyruvate

pyruvate kinase (46, 10, 0, 62)

pyruvate

rank in target list | phase IV drugs | phase III drugs | medicinal chemistry compounds

Arachidonic acid metabolism

arachidonate 5-lipoxygenase (diethylcarbamazine?, filaricide)

arachidonate 5-lipoxygenase (diethylcarbamazine?, filaricide)

leukotriene A4 hydrolase (85, 2, 0, 27)

arachidonate → 5-HPETE → LTA4 → LTB4

prostaglandin G/H synthase (diethylcarbamazine?, filaricide)

PGG2

prostaglandin G/H synthase (diethylcarbamazine?, filaricide)

PGH2

Malate dismutation

oxaloacetate — malate — fumarate — succinate

malate dehydrogenase (29, 0, 0, 0)

fumarate hydratase (67, 0, 0, 0)

succinate dehydrogenase (thiabendazole, pyrvinium, etc.) (40, 0, 0, 0 & 69, 0, 0, 0)

Extended Data Fig. 8: Pathways with known targets.

Suggested novel drug targets (from Supplementary Table 21d) that are in the same biochemical pathway as a known or likely anthelmintic drug target (from Supplementary Table 21b). The known anthelmintic drugs for known or likely targets are shown in red. The rank of the suggested target in our target list (Supplementary Table 21d) is given in parentheses, along with the number (in green) of suggested phase III or phase IV drugs and medicinal chemistry compounds (Supplementary Table 21d).

5