

In the format provided by the authors and unedited.

Comparative genomics of the major parasitic worms

International Helminth Genomes Consortium*

Avril Coghlan^{1*}, Rahul Tyagi², James A. Cotton¹, Nancy Holroyd¹, Bruce A. Rosa², Isheng Jason Tsai^{3,4,5}, Dominik R. Laetsch⁶, Robin N. Beech⁷, Tim A. Day^{8,9}, Kymberlie Hallsworth-Pepin², Huei-Mien Ke³, Tzu-Hao Kuo³, Tracy J. Lee^{3,4,5}, John Martin², Rick M. Maizels¹⁰, Prudence Mutowo¹¹, Philip Ozersky², John Parkinson^{12,13}, Adam J. Reid¹, Neil D. Rawlings¹¹, Diogo M. Ribeiro^{1,14}, Lakshmipuram Seshadri Swapna^{12,13}, Eleanor Stanley¹, David W. Taylor¹⁵, Nicolas J. Wheeler^{8,16}, Mostafa Zamanian¹⁶, Xu Zhang², Fiona Allan¹⁷, Judith E. Allen¹⁸, Kazuhito Asano¹⁹, Simon A. Babayan^{20,21}, Germanus Bah²², Helen Beasley¹, Hayley M. Bennett¹, Stewart A. Bisset²³, Estela Castillo²⁴, Joseph Cook²⁵, Philip J. Cooper^{26,27}, Teresa Cruz-Bustos²⁸, Carmen Cuéllar²⁹, Eileen Devaney²⁰, Stephen R. Doyle¹, Mark L. Eberhard^{30,31}, Aidan Emery¹⁷, Keeseon S. Eom³², John S. Gilleard³³, Daria Gordon¹, Yvonne Harcus^{21,34}, Bhavana Harsha¹, John M. Hawdon³⁵, Dolores E. Hill³⁶, Jane Hodgkinson³⁷, Petr Horák³⁸, Kevin L. Howe¹¹, Thomas Huckvale¹, Martin Kalbe³⁹, Gaganjot Kaur^{6,40}, Taisei Kikuchi⁴¹, Georgios Koutsovoulos^{6,42}, Sujai Kumar⁶, Andrew R. Leach¹¹, Jane Lomax¹, Benjamin Makepeace³⁷, Jacqueline B. Matthews⁴³, Antonio Muro⁴⁴, Noel Michael O'Boyle⁴⁵, Peter D. Olson¹⁷, Antonio Osuna²⁸, Felix Partono⁴⁶, Kenneth Pfarr⁴⁷, Gabriel Rinaldi¹, Pilar Foronda⁴⁸, David Rollinson¹⁷, Mercedes Gomez Samblas²⁸, Hiroshi Sato⁴⁹, Manuela Schnyder⁵⁰, Tomáš Scholz⁵¹, Myriam Shafie¹, Vincent N. Tanya²², Rafael Toledo⁵², Alan Tracey¹, Joseph F. Urban³⁶, Lian-Chen Wang⁵³, Dante Zarlenga³⁶, Mark L. Blaxter⁶, Makedonka Mitreva^{2,54*} and Matthew Berriman^{1*}

¹Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. ²McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA. ³Biodiversity Research Center, Academia Sinica, Taipei, Taiwan. ⁴Department of Life Science, National Taiwan Normal University, Taipei, Taiwan. ⁵Biodiversity Program, Academia Sinica and National Taiwan Normal University, Taipei, Taiwan. ⁶BaNG, Institute of Evolutionary Biology and Centre for Immunity, Infection and Evolution, University of Edinburgh, Edinburgh, UK. ⁷Institute of Parasitology, McGill University, Montreal, Quebec, Canada. ⁸Department of Biomedical Sciences, Iowa State University, Ames, IA, USA. ⁹Institute for 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 Edinburgh, Edinburgh, UK. ¹⁶Department of Pathobiological Sciences, University of Wisconsin-Madison, Madison, WI, USA. ¹⁷Natural History Museum, London, UK. ¹⁸Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. ¹⁹Department of Physiology, Showa University, Tokyo, Japan. ²⁰Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow, UK. ²¹Institute of Immunology and Infection Research, University of Edinburgh, Edinburgh, UK. ²²Institut de Recherche Agricole pour le Développement, Ngaoundéré, Cameroon. ²³Hopkirk Research Institute, AgResearch Ltd, Palmerston North, New Zealand. ²⁴Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay. ²⁵Museum of Southwestern Biology, University of New Mexico, Albuquerque, NM, USA. ²⁶Institute of Infection and Immunity, St. George's, University of London, London, UK. ²⁷Facultad de Ciencias Médicas, de la Salud y la Vida, Universidad Internacional del Ecuador, Quito, Ecuador. ²⁸Department of Parasitology, University of Granada, Granada, Spain. ²⁹Departamento de Parasitología, Universidad Complutense de Madrid, Madrid, Spain. ³⁰Division of Parasitic Diseases and Malaria, Centers for Disease Control and Prevention, Atlanta, GA, USA. ³¹The Carter Center, Atlanta, GA, USA. ³²Department of Parasitology, Chungbuk National University School of Medicine, Cheongju, Korea. ³³Department of Comparative Biology and Experimental Medicine, University of 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

Comparative genomics of the major parasitic worms

International Helminth Genomes Consortium

Supplementary Information

Introduction.....	3
Supplementary Note	3
1. Genomic diversity in parasitic nematodes and platyhelminths.....	3
1.1 Genome sequencing and assembly	3
<i>Sequencing strategy</i>	4
<i>Genome assembly pipeline validation</i>	4
<i>Assembly statistics</i>	5
1.2 Gene-finding and annotation.....	6
<i>Gene-finding, protein names and GO annotations</i>	6
<i>Gene count</i>	6
<i>Assessing gene-finding accuracy using curated gene sets</i>	7
1.3 Variation in genome size.....	8
<i>Evidence for high heterozygosity or recent genome duplication events</i>	8
<i>Effect of repeat content on genome size</i>	10
<i>Effect of coding DNA and intron content on genome size</i>	12
<i>Contribution of repeats, genes, introns and intergenic DNA to genome size variation</i>	13
2. Novel gene families and gene family evolution	14
2.1 Compara database of gene families.....	14
<i>Overview of the Compara database</i>	14
2.2 Species tree, and tree based on gene family presence	15
<i>Phylogenetic pattern of gene family presence and absence</i>	16
2.3 Synapomorphies	17
2.4 Gene family expansions	19
<i>Gene families expanded in parasites</i>	19
<i>Expansions of families related to innate immunity of worms</i>	21
<i>Expansions of gene families involved in development</i>	21
<i>Expansions of gene families involved in host-parasite interactions</i>	22
2.5 SCP/TAPS protein family	26
2.6 GO annotation enrichment.....	27
3. Proteins historically targeted for drug development	28

3.1 Proteases and protease inhibitors	28
<i>Astacins</i>	29
<i>Cathepsin proteases</i>	29
<i>Trypsin proteases, prolyl oligopeptidases, and aminopeptidases</i>	30
<i>Protease inhibitors</i>	31
<i>Trypsin, trypsin-like, and chymotrypsin/elastase protease inhibitors</i>	31
<i>Other protease inhibitors</i>	32
3.2 Novel domain combinations, including those related to proteases	33
3.3 GPCRs	33
3.4 Pentameric ligand-gated ion channels.....	35
3.5 ABC Transporters	36
3.6 The kinomes of nematodes and platyhelminths.....	37
4. Metabolic reconstructions of nematode and platyhelminth parasites	38
4.1 Overall metabolic potential	38
<i>Number of EC numbers and enzymes</i>	38
<i>Variability in KEGG pathway coverage</i>	39
<i>Coverage of KEGG modules</i>	39
4.2 Lipid metabolism.....	40
<i>β-oxidation</i>	40
<i>Ketone bodies</i>	41
4.3 The glyoxylate cycle: linking lipid and carbohydrate metabolism	41
4.4 Carbohydrate metabolism	43
<i>Scavenging carbohydrates from host food</i>	43
<i>Malate dismutation</i>	43
<i>Propionate shunt</i>	44
<i>The lactate dehydrogenase pathway</i>	45
<i>The GABA shunt</i>	45
<i>Glycogen synthesis</i>	46
4.5 Amino acid metabolism	46
<i>Glycine cleavage system</i>	46
<i>Amino acid auxotrophies</i>	47
4.6 Nucleotide metabolism.....	47
<i>Pyrimidine synthesis</i>	47
<i>Purine synthesis</i>	48
4.7 Metabolism of cofactors and vitamins	49
<i>Haem metabolism</i>	49
<i>Vitamin metabolism</i>	49
5. Identifying potential new anthelmintic drug targets and drugs	50
5.1 Summary of approach.....	50
5.2 Ranking helminth proteins as likely drug targets.....	51

5.3 Choice of weightings in the scoring system for targets	52
5.4 Identifying potential new anthelmintic targets, and drugs	52
5.5 Diverse screening set	53
5.6 A high priority target set, giving a smaller diverse screening set	54
Additional references	55
Supplementary Figures	66
List of Supplementary Tables.....	100

Introduction

In this Supplementary Note we provide full details of the methods used for data production at three sequencing centres. In the *Results* section, we also provide data analysis and interpretation at a level of detail that is not possible in the main text. The structure of the document reflects that of the main text, so the underlying evidence and reasoning for each assertion in the main text can be explored in more detail here.

Contributions of individual members of the International Helminth Genomes Consortium are summarised under *Author Contributions* in the main text, and the authors that provided each of the species are identified in Supplementary Table 1. In this document, we have also identified the key authors for each major analysis section.

Supplementary Note

1. Genomic diversity in parasitic nematodes and platyhelminths

1.1 Genome sequencing and assembly

Authors: Avril Coghlan, James Cotton, Kimberlie Hallsworth-Pepin, Nancy Holroyd, John Martin, Philip Ozersky, Isheng Jason Tsai, and Xu Zhang

Sequencing strategy

Many nematode and platyhelminth species are poorly studied and optimised protocols to obtain sufficient high-quality samples for extracting and sequencing DNA do not exist in many cases. Several highly contiguous ‘reference-quality’ genomes were previously published, for which the assemblies underwent laborious manual finishing, including *Schistosoma mansoni*^{10,182}, *Trichuris muris*¹⁷, *Trichinella spiralis*¹⁶, *Strongyloides ratti*¹⁸, and *Onchocerca volvulus*²⁵. Using the N50 as a measure of assembly contiguity (the scaffold size for which 50% of bases are found in scaffolds of this size or longer), these assemblies have N50s in the range 0.4 Mb (*T. muris*) to 32.1 Mb (*S. mansoni*). In contrast, because manual finishing of assemblies is extremely labour-intensive and time consuming, our approach for this project has been to produce draft genomes, and use the reference genomes to aid in gene-finding. As part of the present study, 45 genomes were sequenced (36 at WSI for this project, 6 by MGI and 3 by BaNG, **Supplementary Tables 1 and 2, Supplementary Fig. 26**) and these were combined with 36 published or publicly available genomes, enabling a comparative analysis of 81 nematode and platyhelminth species (**Supplementary Table 2**). The combined set of species spanned each of the major clades of nematodes (clade I, III, IV, V) and flatworms (cestode, trematode), enabling clade-specific genomic changes (e.g. gene family expansions) to be identified.

Genome assembly pipeline validation

To assess the accuracy of the data production methods, the reference *C. elegans* N2 strain was included in the WSI sequencing, assembly and annotation pipelines (**Supplementary Tables 1 and 2**). The resulting assembly was 101.2 Mb, similar to the true size of the N2 genome (102 Mb¹⁸³).

The CEGMA score¹⁰⁹ was 98.3% for full-length matches (99.1% for partial matches), suggesting the assembly included most conserved genes. Furthermore, the WSI gene prediction pipeline (see **Supplementary Note 1.2**) produced 19,282 gene predictions (*c.f.* 20,483 genes in WormBase WS238¹⁵⁴). Taking the longest splice-form for each WormBase WS238 gene, 96% had a BLASTP¹⁰⁷ match (with $E < 1e-5$) to the gene set predicted on our *de novo* assembly, indicating that the assembly was virtually complete, at least with respect to coding DNA.

To check for misassemblies in our draft *C. elegans* assembly, it was aligned to the WormBase reference genome (WS235) using nucmer¹⁸⁴. Since the new assembly was from the same strain as the reference genome, the alignment blocks were filtered to only retain those with $\geq 99.9\%$ identity, and to only keep the best alignment for each region of a scaffold in the new assembly (delta-filter -q option). We then discarded alignment blocks that overlapped by $\geq 90\%$ with an adjacent longer block, resulting in 5,246 alignment blocks that covered 81.2 Mb of the 101.2 Mb new assembly.

To identify interchromosomal misassemblies, the alignment blocks were ordered with respect to their positions along the scaffolds of our new assembly. Considering alignment blocks ≥ 50 kb (453 alignment blocks, which covered 40.3 Mb of our assembly), there were no cases where one scaffold in the draft assembly aligned to two different chromosomes in the reference assembly.

Thus, amongst the long scaffolds there were no obvious interchromosomal misassemblies. However, 254 interchromosomal rearrangement breakpoints were found using scaffolds <50 kb.

To find intrachromosomal rearrangement breakpoints, each scaffold with ≥ 3 alignment blocks to reference chromosomes was considered (since two alignment blocks can only be in the same order in our assembly as in the reference). To find intrachromosomal rearrangements between a particular scaffold and reference chromosome, alignment blocks were numbered along the scaffold (e.g. 1 2 3 4 5) and breakpoints found between stretches of consecutively numbered blocks along the reference chromosome (e.g. two breakpoints: 1 | 5 4 | 2 3). This approach did not consider the orientation of alignment blocks, so inversions spanning just a single block were not detected. However, this method provided an estimate of 93 intrachromosomal rearrangement breakpoints.

Thus, the *C. elegans* assembly produced using the automated pipeline did not contain obvious interchromosomal misassemblies involving large genomic regions, but did contain some interchromosomal misassemblies involving smaller chunks, as well as some intrachromosomal misassemblies. This level of misassembly is relatively low given that this was a draft genome (without long range sequencing and mapping data) and that the primary focus of the present study was on gene content for each species.

Assembly statistics

Contamination scan.

For most species sequenced at WSI, <5% of (bases in) the original assembly was classified as contaminant, but seven species had more contamination: *Brugia timori* (11.6%), *Elaeophora elaphi* (10.3%), *Heligmosomoides bakeri* (7.4%), *Anisakis simplex* (7.3%), *Cylicostephanus goldi* (6.0%), *Syphacia muris* (5.5%), and *Wuchereria bancrofti* (4.6%).

Assembly size and GC content.

For the 45 species sequenced for this project, the assembly sizes varied from 49.1 Mb (*Trichinella nativa*) to 834.6 Mb (*Echinostoma caproni*; **Supplementary Table 2**; **Supplementary Fig. 1**). These were within the range of assembly sizes previously published for nematodes and platyhelminths, from 19.7 Mb (*Pratylenchus coffeae*¹⁸⁵) to 1258.7 Mb (*Spirometra erinaceieuropaei*¹⁴), although free-living platyhelminths such as *Otomesostoma auditivum* have been estimated to have genomes as large as ~18.4 Gb¹⁸⁶. The GC contents of the 45 species were also within the range seen for published species, from 21.3% for *Strongyloides ratti*¹⁸ to 44.5% for *Trichuris muris*¹⁷ (**Supplementary Table 2**).

Genome coverage.

32 out of 45 species had corrected CEGMA¹⁰⁹ scores >85% (see **Methods 1.1**), suggesting their genome assemblies were relatively complete (**Supplementary Table 2**). In 22 cases the corrected CEGMA score was >95%, comparable to those published for many reference genome assemblies that have undergone manual improvement.

Assembly contiguity.

Of the 45 species sequenced for this project, the N50 varied from 1.2 kb to 1.3 Mb (median 25.8 kb, interquartile range 10.1–49.9 kb; **Supplementary Table 2**). However, because genome sizes vary greatly across nematodes and platyhelminths, the ratio of the N50 to the number of scaffolds provides a better measure of assembly contiguity; for example, an assembly with an N50 of 10 kb and 1000 scaffolds is highly contiguous (N50/scaffold-count = 10), while an assembly with an N50 of 10 kb and 100,000 scaffolds (N50/scaffold-count = 0.1) is highly fragmented. Thirty of the 32 genomes with relatively high CEGMA partial scores ($\geq 85\%$) had N50/scaffold-counts of >0.3 . The 13 most fragmented assemblies (N50/scaffold-count < 0.3) had lower CEGMA scores, as expected, suggesting missing genes or that scaffolds were too fragmented for the CEGMA genes to be found (**Supplementary Table 2**).

1.2 Gene-finding and annotation

Authors: Avril Coghlan, Kimberlie Hallsworth-Pepin, Nancy Holroyd, Kevin Howe, John Martin, Philip Ozersky, Eleanor Stanley and Xu Zhang

Gene-finding, protein names and GO annotations

A total of 797,863 genes were predicted for the 45 species (**Supplementary Fig. 27**). The assemblies and gene sets are available via WormBase ParaSite¹⁶⁰.

Taking all predicted proteins for the 45 new genomes, 37.3% were assigned names (**Methods 6**) based on UniProt entries (of which 47.6% were from *C. elegans* orthologues and 12.5% from *S. mansoni*), 22.1% based on InterPro domains, and the remainder were labelled 'Hypothetical protein'. Of the unique names transferred from UniProt (some of which were given to several predicted proteins from different species), 49.6% were transferred from curated ('reviewed') UniProt entries and the other 50.4% from uncurated ('unreviewed') entries.

Gene count

Effect of assembly fragmentation on gene count.

The number of genes per species varied from 9,831 to 37,906 across the 45 species that we sequenced (median 15,892; **Supplementary Table 3**; **Supplementary Fig. 1**). There was a negative correlation between gene count and assembly contiguity, as measured by N50/scaffold-count (Spearman's $\rho = -0.51$, $P = 0.0005$, $n = 45$). In other words, species with fragmented assemblies tended to have higher gene counts, probably because many true genes were split into multiple (partial) gene models. In support of this, there was a strong negative correlation between the gene count and median protein length, showing that high gene counts were due to a higher abundance of split or partial genes (Spearman's $\rho = -0.82$, $P < e-11$, $n = 45$). If this is the case, one would expect each species to have roughly the same amount of coding (exon) DNA, or equivalently, roughly the same total proteome length. Indeed, the total amount of coding (exon) DNA varied far less than the

gene count among the 45 species (interquartile-range/median = 0.27 for total coding DNA, 0.70 for gene count).

Normalised gene counts.

For the most contiguous assemblies, the observed gene count converged to ~9,000-16,000, suggesting the true gene count is in this range for most nematodes and platyhelminths, and that high gene counts of >25,000 were probably artefacts due to assembly fragmentation. To correct for the effect of assembly fragmentation on the gene count, gene counts were normalised for each of our 45 species, by dividing the total proteome length for a species by the mean protein length for *C. elegans* (409.82 amino acids). The normalised gene counts varied from 5,636 to 19,093 (**Supplementary Table 3**). The 5-95% percentile range of the re-estimated counts was 9132-17,274, which agreed well with the 9,000-16,000 estimate for a typical nematode or platyhelminth. The range was however, still lower than the 20,483 genes found in *C. elegans* (WS238). Some of the ~4,500 additional *C. elegans* genes are due to known gene family expansions in the *Caenorhabditis*-lineage (for example of GPCRs¹⁸⁷), but some may be hard-to-find genes that have only been identified as a result of extensive manual curation of the *C. elegans* genome.

Taking just the 33 tier 1 species (those with high-quality assemblies; **Methods 2**), there was no significant difference between the re-estimated gene counts in nematodes (5-95% percentile 10,704-21,142; $n=24$; median 12,426) and platyhelminths (11,215-16,988; $n=9$; median 12,515; Wilcoxon test: $P=0.9$). The four free-living nematodes (all tier 1 species) had higher re-estimated gene counts (median 19,973) than parasitic (tier 1) nematodes (median 12,224; Wilcoxon test: $P=0.01$), suggesting that either there has been gene loss from parasitic lineages or gene gain (e.g. from duplications or birth of novel genes) in free-living lineages. There was just one free-living platyhelminth amongst the 81 species (*Schmidtea*, not a tier 1 species) but its re-estimated gene count (17,096) was relatively high compared to other platyhelminths.

Assessing gene-finding accuracy using curated gene sets

The accuracy of the MAKER gene prediction pipeline used on 36 genomes at WSI (**Supplementary Fig. 27a**) was assessed by using it on the *C. elegans* reference assembly (WormBase WS238), without support from alignments to existing *C. elegans* data. Compared to the highly curated *C. elegans* genes from WormBase, MAKER over-predicted the number of genes (23,902 vs 20,483), and 77% of (unique) WormBase coding exons were predicted with identical coordinates when all coding exons (CDS features) across all splice-forms were considered. The 77% exon specificity obtained using the present WSI MAKER pipeline is lower than could be obtained using gene-finders specifically trained on *C. elegans* data¹⁸⁸. It is, however, a good compromise given our requirement to annotate a wide range of nematode and platyhelminth genomes, without extensive training data for each one.

While accurate prediction of exon boundaries would be ideal, a more practical requirement for gene finding in draft genomes is that the majority of true amino acids are included in the predicted proteins. Even without RNAseq evidence, the WSI MAKER pipeline predicted *C. elegans* genes with 91% sensitivity at the amino acid level (i.e. 91% of true codons were included with the same

coordinates and reading frame in the predicted gene set). To assess the impact on protein annotation (e.g. GO terms, protein names), the Pfam domains from WormBase proteins were compared to those in our predicted *C. elegans* proteins. Of 3,822 unique domains (i.e. unique Pfam identifiers) from WormBase, 96% were found in one or more of our predicted *C. elegans* proteins, showing that MAKER captured most of the expected functional regions.

The accuracy of the WSI MAKER pipeline was further assessed using the assemblies for two published flatworm species, *S. mansoni*¹⁰ and *E. multilocularis*¹³. Compared to partially curated gene models from GeneDB¹⁸⁹, the MAKER pipeline had exon-level sensitivities of 54% and 59% for *S. mansoni* and *E. multilocularis*, respectively. The lower exon-level sensitivity, compared with *C. elegans*, likely reflected the limited curation that these flatworm gene sets have received and the greater difficulty in predicting flatworm genes, which have particularly challenging ‘micro-exons’ and numerous long introns^{13,182}. However, the amino-acid level sensitivities compared well to the sensitivity seen in *C. elegans*, with figures of 82% and 85% for *S. mansoni* and *E. multilocularis*, respectively.

1.3 Variation in genome size

Authors: Avril Coghlan and James Cotton

Among the 81 nematode and platyhelminth species, there was a ~30-fold variation in genome size, from 43 Mb (*Parastrongyloides trichosuri*) to 1259 Mb (*Spirometra erinaceieuropaei*; **Supplementary Table 2**; **Supplementary Fig. 1**). Even between closely related species there was wide variation, for example 341–702 Mb just in the schistosomatid clade of flatworms. Considering the tier 1 species (those with high-quality assemblies; **Methods 2**), platyhelminth genomes were larger than those of nematodes (medians 365 and 100 Mb; $n=9$ and $n=24$; Wilcoxon test: P -value 0.007). Within nematodes, genomes of clade V species tended to be the largest; cestodes usually had similar genome sizes to nematodes (medians 139 Mb versus 100 Mb), but trematode genomes were larger (median 385; Wilcoxon test: P -value 10^{-4} ; **Supplementary Table 2**). There was no difference between the genome sizes of free-living and parasitic (tier 1) nematode species (medians 83 and 102 Mb; Wilcoxon test: P -value 0.3). Some possible reasons for large differences in genome size amongst the nematodes and platyhelminths are: whole (or partial) genome duplication; expansion of repetitive DNA such as transposable elements; increase in intron content; or expansion of gene families. We investigated each of these possibilities in turn below.

Evidence for high heterozygosity or recent genome duplication events

The nematode *Meloidogyne incognita* is hypothesised to be a polyploid that arose due to interspecific hybridisation (i.e. an allopolyploid), and analyses of the *M. incognita* genome revealed many gene loci present in divergent copies^{190,191}. Cytogenetic studies have shown that some other *Meloidogyne* species are diploid or triploid, with variation in ploidy even within species¹⁹². Within flatworms, polyploid strains of species of *Fasciola*, *Paragonimus* and *Schmidtea* have been

observed, and in the case of *Fasciola* and *Paragonimus* may have arisen by interspecific hybridisation¹⁹³.

To investigate whether any of the sequenced helminth genomes revealed traces of recent genome duplication events, we examined the 'low copy' (lc) CEGMA genes that are present as single copy in almost all eukaryotes (CEGMA version 2.4¹⁰⁹). The average number of copies for all of the lc CEGMA genes ('CEGMA (partial) gene count' in **Supplementary Table 2**) should be close to 1.0 unless there has been a whole or partial genome duplication (polyploidisation), large-scale gene duplication has occurred (e.g. in a non-polyploid species), or allelic copies have been (incorrectly) assembled independently (e.g. due to high polymorphism). For example, *Meloidogyne incognita* (downloaded from WormBase release WS245) had an average CEG (partial) number of 1.70, and stood out as a statistical outlier among clade IV nematodes (**Supplementary Table 2**). This approach will only identify relatively recent genome duplication events, since many duplicates that arose from an ancient genome duplication will probably have been subsequently lost.

In clade III nematodes, *Gongylonema pulchrum* was an outlier, with a CEGMA (partial) gene count of 1.53 (**Supplementary Table 2**). However, the assembly for this species was very fragmented (N50/scaffold-count = 0.03), and when shorter scaffolds of <20 kb were excluded, the CEGMA (partial) gene count was 1.09, near the expected value of 1.0. The same effect was observed for the flatworms *Spirometra erinaceieuropaei* (N50/scaffold-count = 0.01; CEG number 1.80; CEG number 1.08 for scaffolds of ≥20 kb) and *H. taeniaeformis* (N50/scaffold-count = 0.42; CEG number 1.48; CEG number 1.07 for scaffolds ≥20 kb). The high CEG number for these three species must be an artefact resulting from many small scaffolds carrying apparently duplicated genes, which in fact represent uncollapsed haplotypes. In contrast, the CEGMA (partial) gene count showed little dependence on scaffold length for *Meloidogyne incognita*, whose genome was probably duplicated (CEG number 1.70; CEG number 1.57 for scaffolds of ≥20 kb). Among clade V nematodes, *Teladorsagia circumcincta* was an outlier with respect to CEGMA (partial) gene count (**Supplementary Table 2**), and CEGMA (partial) gene count showed relatively little dependence on scaffold length (CEG number 1.88; CEG number 1.38 for scaffolds of ≥20 kb). The *T. circumcincta* genome was sequenced based on DNA from multiple individuals, so the high CEG number for this species was probably an artefact due to (incorrectly) uncollapsed haplotypes rather than a genome duplication.

Among the flatworms, *Schmidtea mediterranea* and *Fasciola hepatica* had unusually high CEG numbers (1.98 and 1.42; **Supplementary Table 2**), and these were not dependent on scaffold length (CEG numbers 1.57, 1.32 for scaffolds of ≥20 kb). As mentioned above, polyploid isolates have been observed in these genera¹⁹³. However, the *Schmidtea* genome assembly used (v3.1) is thought to include many uncollapsed haplotypes. In the *F. hepatica* genome publication by Cwiklinski *et al*¹⁹⁴ the authors found no evidence of genome duplication based on CEG number and read coverage, although their assembly was for a different isolate than the one studied here. For a recent genome duplication, large duplicated blocks of genes would be expected. However, looking at paralogues that have arisen since divergence from *Echinostoma caproni* (the closest relative in our data set), no large scaffolds could be found with convincing evidence of genome

duplication (i.e. linked by ≥ 5 pairs of paralogues). Thus, it seems more likely that the higher than expected CEG number reflects many small-scale gene duplications rather than a genome duplication event.

Effect of repeat content on genome size

Finding repeats. There is huge variation in repeat content in published nematode and platyhelminth genomes, varying from $<1\%$ for *Pratylenchus coffeae*¹⁸⁵ to nearly 50% in species such as *Romanomermis culicivorax* (48%¹⁵) and *Schistosoma mansoni* (45%¹⁸²; **Supplementary Table 5**). To annotate transposable elements and other repeat libraries in our assemblies, we built new repeat libraries for each species, and masked the assemblies using RepeatMasker. Of 36 published or publicly available genomes for nematodes and platyhelminths, the repeat content was previously published for 32 species, and despite differences in repeat library construction and assembly versions, for 24 of these 32, our repeat content values were within 5% of the published values (**Supplementary Table 5**). For three species the difference was greater than 10%: *S. mansoni* (published 40%, we found 50.2%), *F. hepatica* (published 55.3%, we found 67.9%), and *C. sinensis* (published 25.9%, we found 38.0%), perhaps because our strategy combined three repeat-finders (RepeatModeler, TransposonPSI, LTRHarvest). The 45 new assemblies contained 2.5–67.9% repeats, exceeding the maximum value of $\sim 50\%$ previously reported for nematodes and platyhelminths (**Supplementary Table 5**). Some of the variation in repeat content may be due to differences in sequencing and assembly approaches; for example, some assemblers discard repeats.

Effect of repeats on assembly size. Among our new assemblies, larger assemblies tended to be more repeat-rich, for example, the repeat-rich flatworms *Trichobilharzia regenti* (59.5%) and *Schistocephalus solidus* (53.8%) had large assemblies (702 and 539 Mb, respectively). However, it is possible that these assembly sizes may underestimate the true genome size. To determine the extent to which this has happened, reads from short-insert sequencing libraries were remapped to the 36 WSI assemblies. Based on read depth, the total amount of extra sequence contained within collapsed repeats was estimated and used to provide new estimates for genome size (see **Methods 1.1**; **Supplementary Table 5**). For three species, $>20\%$ of the original assembly size would be gained by ‘uncollapsing’ those repeats: *Ascaris lumbricoides* (23.9%), *Soboliphyme baturini* (25.8%), and *Schistocephalus solidus* (30.2%). The *A. lumbricoides* (germline) genome contains many copies of satellite repeat that are lost from the DNA of somatic cells by a process of chromatin diminution¹⁹⁵, and these may have been collapsed in the assembly. The collapsed repeats in the clade I nematode *S. baturini* and in the flatworm *S. solidus* may consist of centromeric repeat, as seen in *Trichuris muris*¹⁷, since nematodes in clade I are monocentric¹⁹⁶, as are many flatworms. Alternatively, the collapsed repeats in *S. baturini* — a gonochoristic species with an unknown karyotype — may be part of a Y chromosome, as seen in *Trichuris muris*¹⁷.

Effect of chromatin diminution on assembly. In the early development of clade IIIb nematodes (Ascaridomorpha), the genomes of their somatic cells are reduced by chromatin diminution^{26,59,195,197,198}. In *Ascaris suum*, for example, a 123-bp satellite repeat, comprising about 13% (43 Mb) of the germline genome DNA, is eliminated from somatic cells²⁶. In the present study,

the *A. lumbricoides* DNA from testes (**Supplementary Table 1**) will have largely comprised germline cells (non-diminished genomes) and contained a large amount of collapsed repeat (see above; 76 Mb; **Supplementary Table 5**). The *A. suum* assembly analysed also corresponded to the germline genome²⁶. In contrast, the *Parascaris equorum* and *Toxocara canis* genomes were sequenced from whole individuals and likely corresponded to the somatic (diminished) genome. Thus, repeat content has probably been underestimated in species that undergo chromatin diminution.

Variation in repeat content. We saw huge variation in repeat content across the 81 nematodes and platyhelminths, from 3.8-54.5% (5-95% percentile) when the Ascaridomorpha and *Schmidtea* were excluded since their repeat content estimates were probably inaccurate due to chromatin diminution or heterozygosity, respectively (**Supplementary Table 5**). Even allowing for re-estimated repeat content, within the nematodes, clades III and IV were relatively repeat-poor (90% were $\leq 20\%$ repeat), while clades I and V were repeat-rich (90% were $\geq 19\%$ repeat; **Supplementary Table 5**). The repeat content of cestodes (median 14.2%, range 3.2-53.8%) was similar to that of nematodes (median 18.1%, range 2.8-62.1% excluding Ascaridomorpha), but trematode genomes were far more repeat-rich (median 50.2%, range 38.0-67.9%). Several species that were outliers with respect to genome size were also outliers with respect to repeat content: *Gongylonema pulchrum* in clade III, *Heligmosomoides bakeri* in clade V, and *Fasciola hepatica* among the trematodes (**Supplementary Tables 2, 5**). Massive expansions of particular transposable element families seemed to be responsible for the very high repeat contents of these outliers. *G. pulchrum* had a far higher content of LTR repeats and DNA elements than other members of clade III; and *H. bakeri* had an unusually high amount of LINEs and DNA elements (**Supplementary Table 5**). *F. hepatica* did not have a high content of any known transposon class, but had a very high content of unknown (unclassified) repeats (84.5% of the genome), as did *G. pulchrum*, *S. mediterranea* and *P. xenopodis* (**Supplementary Table 5**). These were possibly transposable elements of unknown class. An analysis of a subset of the nematode genomes studied here concluded that large expansions of transposable element families in particular nematode lineages were due to genetic drift, rather than factors relating to life history or RNAi¹⁹⁹.

Relationship between repeat content and genome size. There was a roughly linear relationship between repeat content and genome size for the 81 nematode and platyhelminth species, suggesting repeat content plays a major role in determining genome size. This was expected, as a positive correlation between repeat content and genome size in eukaryotes has been previously reported⁴⁵. In our data, the correlation was high (Spearman's $\rho=0.96$, $P<10^{-15}$), but the data points were not independent as many species were closely related. In the *C. elegans* genome, there was 83.9 Mb of non-repeat (non-repeat-masked) DNA, and one might imagine that other species have the same amount of non-repeat DNA, and that extra repeat might account for increases in genome sizes relative to *C. elegans*. However, we did not find this to be the case: the amount of non-repetitive DNA also increased with genome size (Spearman's $\rho=0.92$, $P<10^{-15}$). Thus, large genomes may arise in species in which purifying selection is relaxed, in which both expansions of transposable elements and of non-repetitive DNA (e.g. exon or unique non-coding DNA) are tolerated.

Effect of coding DNA and intron content on genome size

Coding DNA content and genome size. There were several species that were outliers with respect to genome size but not repeat content, so other factors such as coding DNA content must play a role in their large genomes: *Romanomermis culicivorax* in clade I; *Ascaris lumbricoides*, *A. suum*, and *Toxocara canis* in clade III, *Globodera pallida* in clade IV; *Teladorsagia circumcincta* in clade V; and *Spirometra erinaceieuropaei*, *Schistocephalus solidus* and *Dibothriocephalus latus* among cestodes (**Supplementary Tables 2 and 5**). As mentioned above, the repeat content and genome size estimates for the Ascaridomorpha were probably inaccurate due to chromatin diminution. Of the others, *R. culicivorax*, *S. erinaceieuropaei*, and *S. solidus* were all outliers with respect to coding DNA content in their respective clades (**Supplementary Table 3**). All of these had unusually high gene counts compared to other members of their clade (after correcting for assembly fragmentation; **Supplementary Table 3**). Their high gene count seemed to be partly due to gene duplications, as the number of paralogues was particularly high in those species (**Supplementary Table 3**). Our CEGMA analysis (see *Evidence for high heterozygosity or recent genome duplication events* above) showed that although *S. erinaceieuropaei* had a high CEG number, it appeared to be an artefact resulting from many small scaffolds carrying apparently duplicated genes, which were actually uncollapsed haplotypes. Like *S. erinaceieuropaei*, *R. culicivorax* and *S. solidus* both had a relatively high level of assembly fragmentation (all had N50/scaffold-count ≤ 0.6), so possibly also suffered from this issue. However, *R. culicivorax* had a greater number of gene families that contained *C. elegans* members compared to other clade I nematodes (4,728 in *Romanomermis* versus 3,685–4,155 in other clade I species), so may have a larger gene set because it has retained ancestral nematode genes that other clade I species have lost.

Intron Content and Genome Size. An unusually large amount of intron DNA compared to other members of their clades seemed to have contributed to the large genome sizes of *Romanomermis culicivorax*, *Ascaris lumbricoides*, *A. suum*, *Toxocara canis*, *Teladorsagia circumcincta* and *Schistocephalus solidus* (**Supplementary Table 3**). This may partly be due to an unusually high total number of introns in *Romanomermis culicivorax*, *Ascaris lumbricoides* and *A. suum* compared to other members of the same clade, probably due to their high gene counts (**Supplementary Table 3**). High gene counts may be due to gene loss from other clade I species in the case of *R. culicivorax* (see above), and due to diminution in Ascaridomorpha. These species also had unusually long median intron size compared to other members of the same clades (medians 1,345 bp, 979 bp, 1,031 bp; **Supplementary Table 3**). In addition, *T. canis* also had unusually long introns (median 738 bp), as did *Teladorsagia circumcincta* and *Schistocephalus solidus* (medians 340 bp, 1,264 bp). There could be various reasons for their high median intron size: expansion in size of existing introns (whose positions in genes are conserved with other species), or gain of novel long introns, or loss of short introns by these species; or alternatively, due to shrinking of conserved introns, gain of short introns or loss of long introns by other species.

Contribution of repeats, genes, introns and intergenic DNA to genome size variation

Multiple regression model of genome size. As many of the factors that were correlated with genome size were themselves correlated with each other, we built a statistical model to identify the most important factors driving genome size variation in these data. As an initial (complete model) we began with a set of covariates that seemed likely to be independent of each other, including variates designed to capture variation in gene content (number of genes, total exon length and total intron length), the degree of gene duplication (number of paralogues), the quality of assembly (N50n: the number of scaffolds whose summed length is N50) and the repeat content (total SINE length, total LINE length, total length of LTR elements, total length of DNA elements, total length of small RNA genes, total satellite repeat length, total simple repeat length and total length of low-complexity sequence). GC content would not be expected to affect genome size and was included as a control. Stepwise regression was used to identify a minimal model that best explained the data, by adding or removing terms at each step that had the largest effect on the Akaike Information Criterion. The resulting model contained eight terms including the intercept, all of which were significant at $P < 0.01$, except for the total length of small RNA genes. The latter was subsequently dropped to leave a minimum adequate model of six terms (in order of increasing P -value: total LTR length, total length of simple repeats, N50n, total length of DNA elements, total length of introns, total length of low complexity sequence) plus a significant intercept term.

The non-repetitive intergenic length was omitted from the regression above, as this is simply the remainder of sequence not annotated elsewhere. Indeed, a regression model including total intron DNA, total repetitive intergenic DNA, total non-repetitive intergenic DNA and total coding DNA is essentially saturated (residual deviance=0.004, due to some figures being rounded to kb). However, non-repetitive intergenic length is correlated with total genome size ($\rho = 0.846$). Furthermore, a model dropping the non-repetitive intergenic DNA term had residual deviance 328,646 confirming that this term explains a significant amount of variance ($p \sim 0$ by chi-squared test with 77 degrees of freedom).

Bayesian mixed-effect model of genome size. A complication with the regression model was that the species varied in their degree of relatedness; for example, *Strongyloides* was represented by multiple species but species such as *Romanomermis*, *Dracunculus* and *Protopolystoma* were the only exemplars of large parasite groups and were only distantly related to other species. This effectively led to non-independence of the species. To address this, a phylogenetic mixed-effect model was built using the R package *MCMCglmm*¹³⁴, in which the species tree was used to specify the expected covariance between species and so correct for this non-independence. Including the phylogenetic effect showed similar results to the multiple regression model, with the phylogenetic component only explaining approximately 9% of the variance in genome size, although this estimate had very wide confidence intervals (0.02%-49%). **Supplementary Table 6** shows the coefficient estimates, and MCMC P -values for tests that the coefficient estimates were significantly different from zero.

These models confirmed that repeat content, particularly some transposable elements (DNA elements and LTR-containing elements) and simple repeats, explained much of the variation in

genome size. There was also an effect of assembly quality (N50n), but the number and length of coding regions did not appear to play a significant role. Surprisingly, a high content of low complexity DNA was significantly associated with smaller genome size; clade III and IV nematodes had a relatively high fraction of low complexity DNA (although still only 2-3% of the genome; **Supplementary Table 5**) but relatively small genomes (**Supplementary Table 2**). Genome size variation in parasitic nematodes and platyhelminths appeared to be largely due to changes in a number of different non-coding elements, 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

2.1 *Compara database of gene families*

Authors: Avril Coghlan, James Cotton, Bhavana Harsha and Eleanor Stanley

Overview of the Compara database

Due to splits or partial gene predictions in the more fragmented assemblies, high observed gene counts may not represent true increases in gene number. To get a more accurate picture of which gene families have expanded in different clades, a database of gene families was constructed using the Ensembl Compara pipeline⁴⁶ based on sequence similarity. The database included 91 species: 56 nematodes (four free-living), 25 platyhelminths (one free-living), and 10 outgroup species from seven further animal phyla (Porifera, Annelida, Chordata, Cnidaria, Mollusca, Arthropoda, Placozoa; **Supplementary Table 2**).

The Compara pipeline splits up families of >750 genes; for such cases, we merged these subfamilies back into one large family, resulting in 109,571 families. Thus, from the combined data set of 1,639,427 genes (91 species), 109,571 families were created, comprising 1,365,243 genes. A phylogenetic tree was built for each family and orthologues and paralogues were identified. From this list, 1,220 gene families were identified that contain genes with Pfam domains that are associated with transposable elements, and these were excluded from further analysis, leaving 108,351 families containing 1,345,329 genes (**Supplementary Data File 1, Supplementary Table 7**). Most families (66%) have 2-5 genes while 2,107 families have >100 genes, the largest family containing 1219 genes.

Nematode and platyhelminth families that are missing from model organisms. Among the 108,351 families, there are 33,593 families that have ≥ 2 nematode species but lack *C. elegans* and 20,304 families that have ≥ 2 flatworm species but lack *S. mansoni*, indicating that there is a considerable breadth of nematode and platyhelminth gene functions that are absent from these model organisms.

Hypothetical proteins. The mean percentage of genes to which we could not assign a protein name (see **Methods 6**) per species was 37.0% across the 81 helminths (5-95% range 24.3–56.2%).

Many of these ‘hypothetical proteins’ are likely to represent false positive gene predictions, for which conservation across species would not be expected. However, when we classified Compara families as ‘hypothetical’ if $\geq 90\%$ of the genes in the family had the product description ‘hypothetical protein’, 51,356 (46.9%) families were found to be ‘hypothetical’. That is, although only one third of genes are annotated as hypothetical, nearly half of families contain mostly hypothetical proteins. This is at least partly explained by the finding that hypothetical families tend to be smaller ($P < 2.2\text{e-}16$; one-sided Kolmogorov-Sminov test; **Supplementary Fig. 4a**). There are two possible reasons for this: that some hypothetical families represent small collections of spurious gene models and that small (probably clade-specific) families are less well studied. While most are small, 5,260 hypothetical families have at least 10 members and 2,875 are found in at least 10 species (**Supplementary Fig. 4b**).

2.2 Species tree, and tree based on gene family presence

Author: James Cotton

A whole-genome maximum likelihood phylogenetic tree for helminths and ten members of other animal phyla was constructed based on 202 gene families from our Compara database that were present in at least 25% of species, and were single-copy in every species in which they were found (**Fig. 1**, **Supplementary Fig. 3**). To quantify the degree to which gene family content reflects the phylogeny, we also constructed a maximum-likelihood phylogeny based on gene family presence and absence for families that are not shared by all 81 species. This ‘family sharing’ phylogeny supports 62 out of 89 internal nodes in the amino acid alignment species tree.

In a tree based on ribosomal RNA, the Spiruromorpha+other (clade IIIc) grouped with all but two Ascaridomorpha (clade IIIb), with Oxyuridomorpha (clade IIIa) as an outgroup, albeit without bootstrap support²⁰⁰. A second ribosomal RNA tree also agreed with this²⁰¹. Our data set includes the first two genomes for oxyuridomorphs, making it possible to investigate this question. Consistent with the rRNA, our sequence tree supports a grouping of Spiruromorpha+other and Ascaridomorpha, with Oxyuridomorpha as an outgroup (bootstrap 92%; **Fig. 1**). In addition, the tree based on gene family presence/absence groups Spiruromorpha+other with Ascaridomorpha, with Oxyuridomorpha as an outgroup (bootstrap 90%, **Supplementary Fig. 3**). Oxyuridomorphs and Ascaridomorphs do not use intermediate hosts, while many Spiruromorphs use vector hosts to transit from one definitive host to another²⁰², so Ascaridomorpha and Oxyuridomorpha may have independently lost use of an intermediate host.

Another interesting question is the placement of the genus *Bursaphelenchus*, which is grouped with *Panagrellus* in our tree (bootstrap 76%), as a sister group to *Meloidogyne* and *Globodera* (**Fig. 1**), and this is also supported by our tree based on gene family presence/absence (**Supplementary Fig. 3**, bootstrap 100%). Our tree is also consistent with a published rRNA tree, which grouped *Bursaphelenchus* with *Panagrellus*, with *Strongyloides* as an outgroup, although without bootstrap support and although the authors pointed out that GC content variation may bias the rRNA tree²⁰⁰. In contrast to our result, *Bursaphelenchus*, *Meloidogyne* and *Globodera* have

previously been classified in the Tylenchomorpha, while the superfamily Panagrolaimoidea, to which *Panagrellus* belongs, was in the infraorder Panagrolaimomorpha, along with the *Strongyloides* species and other clade IVa species²⁰³ (**Supplementary Table 4**). While bootstrap support for the placement of *Bursaphelenchus* with *Panagrellus* is not complete (76%), our analysis adds to data suggesting that the Tylenchomorpha may not be monophyletic.

The lungworm *Dictyocaulus* is grouped with the lungworm genus *Angiostrongylus* in our tree (with 100% bootstrap support; **Fig. 1**) and also in our tree based on gene family presence/absence (69%; **Supplementary Fig. 3**), in agreement with a previously published rRNA tree (bootstrap 57%²⁰⁰). This is surprising because *Angiostrongylus* is in the family Metastrongylidae, and *Dictyocaulus* is in Trichostrongylidae (**Supplementary Table 4**) but is not placed in our tree with other Trichostrongylidae such as *Haemonchus* and *Nippostrongylus* (clade Va; strongylid, other). This suggests that the family Trichostrongylidae is not monophyletic.

Previous analyses of most single-gene data sets, and of complete mitochondrial genome data, supported either monophyly of cestodes and trematodes with monogeneans forming an outgroup to these two groups, or supported a clade of monogeneans and cestodes. The former arrangement suggests that complex, endoparasitic life-cycles evolved only once in Neodermatan platyhelminths (monogeneans, trematodes and cestodes)²⁰⁴, while the latter is traditional and supported by some morphological similarity between the hooked haptors of monogeneans and the onchospherical hooks of tapeworm larvae²⁰⁵. In our species tree, the monogenean *Protopolystoma* is grouped with the trematodes (bootstrap 90%; **Fig. 1**), but this relationship is not recovered in the tree based on gene family presence/absence (**Supplementary Fig. 3**); our gene content analysis shows *Protopolystoma* as sister group to both cestodes and trematodes. The only previous analysis of a genome-wide data set for a monogenean parasite (*Gyrodactylus salaris*) also recovers this relationship and conflicts with our sequence-based result²⁰⁶, although this is of a very distantly related monogenean, and it is not clear that Monogenea is a monophyletic group^{207,208}.

Disagreements between our species tree, and the tree based on gene family presence/absence, are mostly on relationships within and between the outgroups, and within closely-related groups of species such as the genera *Trichuris*, *Hymenolepis* and *Schistosoma* (**Supplementary Fig. 3**). The concatenated sequence data and gene content analyses are likely to be affected by different biological processes beyond shared ancestry, and the gene content data probably has less power to resolve some relationships.

Phylogenetic pattern of gene family presence and absence.

As expected, the pattern of gene family sharing between the 81 nematodes and platyhelminths largely reflects the phylogeny of these organisms. Particularly striking is the very large number of families specific to one or few species (**Supplementary Fig. 2a**): these are almost entirely small families with just one or a few genes per species (**Supplementary Fig. 2b**). The majority of families (76%) include ≤ 5 species, and most of these families with ≤ 5 species (86%) have ≤ 5 genes in total. Larger families are shared between the highest-level taxonomic groups (**Supplementary Fig. 2a**): for example, 1665 gene families are shared by clade III and V

nematodes but missing from clade IV, consistent with our species tree (**Fig. 1**). Other families (1,545) are shared by clades III, IV and V, and many families are shared by all nematodes and some outgroup species (1,556), but far fewer families (508) are nematode-specific and shared by the four nematode clades, confirming the very different gene content between the clade I nematodes and other nematode groups included here. There are also many gene families specific to Neodermata (*Protopolystoma*, trematodes and cestodes; 965) and there are many more (1,329) exclusive to just the endoparasitic groups (trematodes and cestodes). While this is intriguing biologically, it does not appear to reflect the arrangement proposed in our species tree, where *Protopolystoma* is placed as sister to the trematodes, as rather fewer (725 families) are specific to trematodes+*Protopolystoma*. Strikingly, few gene families (just 233) are unique to and shared by all the flatworms here, and the free-living *Schmidtea* has more families shared with outgroup species than with other platyhelminths.

2.3 Synapomorphies

Authors: Mark Blaxter and Dominik R. Laetsch

We analysed the set of Compara gene families for families which constituted synapomorphies at key nodes relevant to parasitism within the phylogenetic tree of Metazoa (see **Methods 7**). In total, 3,512 synapomorphic Compara families displayed a taxon coverage of 100% (i.e. 'complete' synapomorphies), of which 16.9% could be assigned a representative functional (InterPro) domain annotation (**Supplementary Table 8**). Allowing for stochastic absence of species in gene families (defining 90–100% taxon coverage as 'stochastic absent' synapomorphies), 2,369 additional synapomorphic families could be identified, which displayed a higher percentage (24.0%) of representative functional domain annotation. The low percentage of representative functional domain annotations arises from the fact that the sequences in these families are highly diverged and in many cases do not contain members from well-studied organisms. However, it is striking that many phylogenetically conserved gene families at deep nodes have escaped functional annotation.

For the two phyla, nematodes and platyhelminths, we found 124 and 142 synapomorphic families, respectively, of which roughly half (51.6% for nematodes and 64.8% for platyhelminths) received a representative functional annotation (**Supplementary Table 8**). Families with complete taxon coverage across phyla were less frequent in nematodes (20) than in platyhelminths (71).

Between the two phyla, clear differences between nematodes and platyhelminths in functional annotation of synapomorphies could be observed for those related to cellular structure such as cell adhesion, membrane and microtubules (**Fig. 2**). Nodes within platyhelminths exhibited higher numbers of synapomorphies related to cellular structure such as cell adhesion (46 synapomorphies), membrane (28), and microtubule (13) than within nematodes (3, 21, and 4, respectively). Furthermore, within platyhelminths the 'births' of these families seemed to have occurred in a more gradual and balanced manner along both major lineages, Cestoda and Trematoda. In nematodes, innovations related to membrane structure seemed to have occurred

primarily at the last common ancestor of all nematodes, with piecemeal emergence of novelty along nodes of clade I, clade IVa, and clade V (**Fig. 2**).

Within Neodermata (*Protopolystoma*, trematodes and cestodes), a clade-specific inositol-pentakisphosphate 2-kinase (EC 2.7.1.158, family 922204) was identified which might be involved in storage of phosphates in the form of calcium-hexakisphosphate, as has been shown for the laminated layer of larval *Echinococcus*⁵⁰.

Amongst nematodes, both clade III (one synapomorphy) and clade IV (5 synapomorphies) displayed independent synapomorphies related to fatty acid retinoid (FAR) binding, distinct from the Compara family containing Gp-FAR-1 described from *Globodera pallida*⁶⁶. FARs have been previously suggested as vaccine targets²⁰⁹, since many parasitic nematodes rely on importing essential lipids from their host.

Ferrochelatase is an enzyme that catalyses the terminal step of haem biosynthesis. Nematodes encode ferrochelatase-like (FeCL) proteins that are devoid of the active site, while certain onchocercid nematodes (*B. malayi*, *A. viteae*, *D. immitis*, and *O. volvulus*)⁴⁸ and *S. venezuelensis*⁴⁹ harbour a second, apparently functional, copy of ferrochelatase acquired through horizontal gene transfer (HGT) from α -proteobacteria. The Compara family containing these functional HGT-ferrochelatases (family 787620) included taxa from clade III (with the exception of *A. simplex*, *A. suum*, *A. lumbricoides*, and *P. equorum*) and IV (with the exception of the three plant-parasitic species). This family also included two platyhelminths (*S. mediterranea* and *T. regenti*), two arthropods (*D. melanogaster* and *I. scapularis*), the sponge (*A. queenslandica*) and *T. adhaerens*.

Functional ferrochelatases (containing both the Ferrochelatase domain, IPR001015, and the Ferrochelatase active site domain, IPR019772) were found in the Compara family 787620 mentioned above, composed mainly of nematode sequences, and family 850580, harbouring sequences from non-nematode taxa. Ferrochelatase-like sequences (devoid of active site) were recovered from family 740872 and family 1184543. Phylogenetic analysis of the ferrochelatase Compara families 787620, 1184543, 850580, 740872, and bacterial ferrochelatases retrieved from NCBI (see **Methods 8**) placed the additional, non-nematode ferrochelatases (with the exception of *S. mediterranea*) together with those of family 850580 (composed solely of platyhelminth and outgroup proteins) (**Supplementary Fig. 5**), suggesting clustering of these with the HGT-ferrochelatases was simply an artefact of the Compara analysis. Sequence similarity searches of the scaffold from which the *S. mediterranea* protein originated suggested it was a contaminant from an α -proteobacterium. The phylogenetic analysis also suggested that the acquisition of functional ferrochelatase by HGT predates the split of clade III and IV and that it was subsequently lost in clade V (**Supplementary Fig. 5; Fig. 1**).

We also found a synapomorphic family (family 1184543) of clade IVa, annotated as ferrochelatase but lacking the active site. Sequences from this family were placed as sister to nematode FeCL orthologues (**Supplementary Fig. 5**). Currently, the functional role of these proteins is unknown⁴⁸.

2.4 Gene family expansions

Authors: Avril Coghlan, James Cotton, Stephen Doyle, Nancy Holroyd, Taisei Kikuchi, Adam Reid, Diogo Ribeiro, Bhavana Harsha, Gabriel Rinaldi, Bruce A. Rosa and Rahul Tyagi

To control for differences in gene counts due to assembly fragmentation (e.g. genes split across multiple contigs) in our analysis of gene family expansions, we used summed protein length per species for each family as a proxy for gene counts; and only 'tier 1' species (those with highest quality assemblies) were considered when creating our initial list of candidate expansions (see **Methods 9**), although all species were considered during subsequent manual curation of each expanded family.

Three separate scoring metrics and two clade definitions were used to identify gene families that were specific to, or expanded in, different clades and that therefore may underlie clade-specific traits. These metrics were calculated for each of the 108,351 families (after excluding 1220 transposon families; see Protocol Exchange¹⁴⁹) in our in-house Ensembl Compara database. In very small families, the estimates of the three metrics C_v , Z_{max} and E_{max} were very noisy (because the underlying estimates of mean and standard deviation were noisy), so we discarded families having <10 genes from tier 1 species (those with high-quality assemblies; **Methods 2**), leaving 10,986 families. Each metric was calculated for both 'narrower species groups' (e.g. 'Trematodes - Schistosomatids'; see 'Analysis group' in **Supplementary Table 4**) and 'broader species groups' (e.g. 'Trematodes'). Because the coefficient of variation is independent of the species (it is calculated based on all genes from all species in the gene family), for each family we calculated five variables reflecting variability: C_v , Z_{max} and E_{max} for the original species groups, and Z_{max} and E_{max} for the broader species groups.

We filtered the list of 10,986 families to identify those with the most striking patterns of variability. Taking the union of the top 500 highest-scoring families according each of five metrics (C_v , Z_{max} and E_{max} for the narrower species groups, and Z_{max} and E_{max} for the broader species groups), yielded 1,248 families (**Supplementary Table 9**). Based on Pfam²¹⁰ domain content, BLAST matches, and unusual sporadic distribution across the species tree, 58 of these 1,248 families were classified as likely transposable elements (**Supplementary Table 9b**). Four of these contained *C. elegans* genes with viral or transposable element origins. We discarded the 58 putative transposable element families from our analysis, as well as an additional 30 families in which the gene count was as high (or nearly so) in outgroups as in nematodes and platyhelminths (**Supplementary Table 9b**). We also identified 165 families that were expanded in free-living nematode and platyhelminth species (**Supplementary Table 9b**).

Gene families expanded in parasites

Previously reported expansions.

The remaining 995 families had expanded in parasitic species. These included 25 gene family expansions previously reported, 21 of which have proven or hypothesised links to parasitism (**Supplementary Table 9a**).

Kinase, protease, protease inhibitor and GPCR expansions.

Expansions in parasitic nematodes and platyhelminths were found in 43 families annotated as proteases (8) and protease inhibitors (8), GPCRs (12), and kinases (15) (**Supplementary Table 9a**). Genes in these key functional classes were investigated in further detail in **Supplementary Note 3.1, 3.3, 3.6** respectively.

Other gene family expansions identified by manual curation.

For the remaining 927 families, the distributions of gene counts per species were manually inspected across the species tree. We identified 176 families (described below) with convincing expansions across many species, or large expansions that were present in a couple of parasitic species (**Supplementary Table 9a**). Since assembly and gene-finding artefacts can create apparent species-specific expansions, particularly from polymorphic loci, single species expansions of < 20 genes were ignored in most cases, except for those in nematode clades I and IIc where fewer expansions were seen overall.

Several striking expansions involved core biological processes that cannot easily be attributed to particular parasitic adaptations. For instance, family 451451 is particularly expanded in hookworms but encodes ribosomal proteins. Similarly, we identified an expansion of a Xeroderma pigmentosum group G (XPG) family, particularly in schistosomatids (family 437828; 9 *S. mansoni*, 9 *S. haematobium*, 4 *Clonorchis*; 3 *F. hepatica*; **Supplementary Fig. 29** and **Supplementary Table 9a**) that is likely involved in nucleotide excision repair of DNA²¹¹⁻²¹³. It is tempting to speculate that expansion of this family might enable repair of DNA damage that results from the rapid asexual replication of sporocyst stage parasites within a snail host. Another possibility is that this family may be involved in repair of DNA damage incurred during the free-living stages of their life cycle (including cercariae, which are exposed to UV light prior to finding a host)²¹⁴⁻²¹⁶.

Among the 176 families that we identified as expanded in parasitic helminths (see *Gene family expansions* in **Supplementary Note 2.4**), 32 comprised hypothetical proteins (**Supplementary Table 9a**; **Supplementary Fig. 7**). Most groups of parasites possessed families in this list, suggesting a wide range of new biology is waiting to be uncovered. However, some of these may be difficult-to-identify transposon-related sequence families, something we have seen for many highly variable families. Most of the ‘hypothetical’ families lacked signal peptides and transmembrane domains with the exception of the 96-member family 393312 that contained 25 proteins with predicted signal peptides (but no transmembrane domains), suggesting a secretory role. The family is strikingly expanded in clade Va, particularly *Teladorsagia* (26), *Haemonchus* (22) and *Oesophagostomum* (13) (**Supplementary Fig. 7g**).

In addition to expansions of uncharacterised families, 114 large expansions in poorly characterised gene families were observed (9 of which were particularly striking; **Supplementary Table 9a**). For example, in trematodes there was a striking expansion of a sulfotransferase gene family containing *S. mansoni* gene *Smp_089320*, which activates the pro-drug oxamniquine that is used to treat schistosomiasis (family 219111; 9-16 copies in trematodes versus 1-6 in cestodes; **Supplementary Fig. 7j**). Sulfonation of oxamniquine by *Smp_089320* produces a reactive product capable of alkylating DNA, proteins and other macromolecules⁵³; however, the native function of

Smp_089320 and its orthologues in other trematodes (which is presumably unrelated to oxamniquine) is not known.

Expansions of families related to innate immunity of worms

Like free-living species, parasites encounter other pathogens such as bacteria, fungi and viruses both within the host and during free-living stages, and genes involved in innate immunity are likely to be under strong selection. Glycosyl transferases may contribute to resistance of nematodes to microbial pathogens, by allowing them to continually modify their surface coat²¹⁷. In clade IVa nematodes, we saw expansion of a GT31 galactosyltransferase family that included *C. elegans bus-4*, which was implicated in protection against bacterial infections⁵⁴ (family 47681; EC 2.4.1.122, 18-37 copies per species; **Supplementary Fig. 8a** and **Supplementary Table 9a**), although it is of course possible that genes in this family have diverse functions. We also saw expansion in clade IVa of a family containing the *C. elegans* infection response gene *irg-3*, which is induced in *C. elegans* upon infection with pathogenic *Pseudomonas aeruginosa* (see Table 2 in Shapira *et al*⁵⁵; family 781277; 75% predicted secreted; 3-8 per *Strongyloides* species; **Supplementary Fig. 8b**).

In nematodes clades Va and Vc, we observed an expansion of a lysozyme (GH22) gene family (family 95037; 68% predicted secreted; 19-31 copies in *Ancylostoma* species, 7-13 in *Haemonchus* species; **Supplementary Fig. 8c** and **Supplementary Table 9a**). Lysozyme damages bacterial cell walls by degrading the glycosidic bond linking *N*-acetylglucosamine and *N*-acetylmuramic acid in the proteoglycan of cell walls⁵⁶, and the *C. elegans* genes in this family (*lys-1*, *lys-2*, *lys-3*, *lys-7*, *lys-8*, *lys-9*) have been implicated in resistance to infection by certain bacteria²¹⁸⁻²²¹. The excretory/secretory products of *Ancylostoma caninum* and *Heligmosomoides bakeri* include lysozyme-like proteins similar to *lys-1*, *lys-2*, *lys-3*, *lys-7* and *lys-8*²²², and Hewitson *et al*⁵⁶ suggested that these lysozymes modify the bacterial populations cohabiting with the parasites in the host intestine. In clades Va and Vc we also saw a small expansion of a haem peroxidase family containing *C. elegans bli-3* and *duox-2*, both of which are involved in collagen/cuticle biosynthesis wherein they function in dityrosine cross-linking of collagen, and so are important for cuticular integrity; but *bli-3* has also been implicated in generation of reactive oxygen species as part of an innate immunity mechanism that protects against fungal and bacterial infection^{57,223} (family 184998; EC 1.11.1.7/1.6.3.1; 3–7 per *Haemonchus* species, 10 *T. circumcincta*, 7 *A. caninum*, 6 *S. vulgaris*, 6 *O. dentatum*; **Supplementary Fig. 8d** and **Supplementary Table 9a**).

Expansions of gene families involved in development.

Chromatin diminution is a striking aspect of development in clade IIIb nematodes whereby regions of the genome are deleted in early embryonic cells, during each cell division, except in cells set aside to become future gonads. The phenomenon has been observed in *A. suum*, *A. lumbricoides*¹⁹⁵, *P. equorum*¹⁹⁸, and *T. canis*²²⁴. Accordingly, we saw an expansion in these clade IIIb nematodes of a gene family that contained orthologues of the *Parascaris univalens* coiled-coiled protein PUMA (family 114493; 34% predicted secreted; 16–40 per *Ascaris* species; 18 *P.*

equorum, 25 *T. canis*, 34 *A. simplex*; **Fig. 2b** and **Supplementary Table 9a**). During mitosis in *Parascaris*, PUMA is localised to the centrosomes, kinetochore-microtubules, and continuous centromeric region of the holocentric chromosomes⁵⁸, but its molecular function is unknown. The genes in this family had the Rootletin Pfam domain (PF15035), so-called after the human CROCC (Rootletin) protein that is required for centrosome cohesion before mitosis²²⁵. The one *C. elegans* gene in the family, *Ifi-1*, is expressed in an area surrounding the kinetochore-microtubules during metaphase²²⁶. We conjecture that in *C. elegans* and most nematodes this gene family plays a role in kinetochore and/or kinetochore-microtubule function, and that expansion of this family in clade IIIb is somehow linked to chromatin diminution, since segments of chromatin that are to be eliminated lack any kinetochore¹⁹⁸. However, it was also expanded in *Anisakis*, not known to undergo diminution.

Helminths undergo developmental changes such as moulting throughout their life cycle, including some specific to certain clades. For example, L1s of clade IVa species can either develop in a 'direct' route via an infective L3, or an 'indirect' route via a free-living L3¹⁸, while *Bursaphelenchus xylophilus* is the only parasitic nematode known to form dauers^{19,227}. In clade IVa, and in *Bursaphelenchus*, we saw expansion of a family with the 'ecdysteroid kinase' Pfam domain (family 147797; 7-20 per *Strongyloides* species; 34 *Bursaphelenchus*; **Supplementary Fig. 8e** and **Supplementary Table 9a**). This family was absent from *C. elegans*. In insects, ecdysteroids play a central role in controlling the onset of moulting, metamorphosis and reproduction²²⁸, and after their synthesis, insect ecdysteroids are phosphorylated by ecdysteroid kinases to an inactive form²²⁹. Ecdysteroids may play a role in regulating developmental events such as moulting in some (but probably not all) nematodes²³⁰, and some (but not all) have orthologues of ecdysteroid receptors²³¹⁻²³⁴. This ecdysteroid kinase Compara family may possibly regulate the activity of other steroid hormones in clade IVa and *Bursaphelenchus*, such as the dafachronic acids that regulate dauer entry in *Bursaphelenchus*²³⁵ and the direct/indirect developmental switch in *Strongyloides stercoralis*⁶⁰. Expansion of this family may have been involved in evolution of these developmental choices.

Expansions of gene families involved in host-parasite interactions

Infections with parasitic worms are typified by their chronicity. Worms have evolved mechanisms to evade or modulate their host's immune response, enabling infections to be established within specific host niches, where infections with individual worms can last years or decades²³⁶. Amongst many of the convincing parasite expansions were numerous examples with plausible links to host-pathogen interactions.

In strongylid nematodes, there was a striking expansion of an apyrase (EC 3.6.1.5) gene family that included *C. elegans* *apy-1* (family 135728; 34% predicted secreted; 15-24 copies per species in *Ancylostoma*; 8-10 in *Haemonchus*; **Fig. 2b** and **Supplementary Table 9a**). Release of ATP by host cells is a 'Danger-Associated Molecular Pattern' (DAMP), signalling cell or tissue injury. An apyrase related to *apy-1* is secreted by some haematophagous insects and dampens this host 'danger' signal by hydrolysing host ATP^{237,238}. Nisbet *et al*²³⁹ suggested that a *Teladorsagia* member of this family that is present in excretory/secretory products of L4 larvae (as is a

Heligmosomoides member²⁴⁰) and hydrolyses ADP/ATP, may play a similar immunomodulatory role. In addition, since many strongylids are blood feeders (**Supplementary Table 12**), apyrase may prevent purinergic activation of platelets, similar to its role in haematophagous insects, thus preventing clotting and permitting blood feeding. Interestingly, one clade of this Compara family had a novel Pfam domain combination, with both apyrase (Pfam:PF06079) and flavin-containing amine oxidoreductase (Pfam:PF01593, monoamine oxidase-like) domains (e.g. *H. contortus* HCOI00605900, *N. americanus* NECAME_12085, *A. caninum* ANCCAN_11441). Because proteins with the amine oxidoreductase domain likely break down amines²⁴¹, strongylid genes with this domain combination may modulate the levels of host amines, such as histamine, another ‘danger’ signal indicating tissue damage^{62,242}. We also found an independent smaller expansion of another apyrase gene family, containing human genes *ENTPD1*, *ENTPD2*, *ENTPD3*, *ENTPD8*, in the cestode genus *Taenia* (family 511981; 32% predicted secreted; 9-12 copies per *Taenia* species; 1-2 per *Hymenolepis* species; **Supplementary Fig. 9a** and **Supplementary Table 9a**). It was previously shown that the tegumental membrane of *Taenia crassiceps* cysterci has an apyrase, which has been hypothesised to degrade nucleotides in the bloodstream of the host in order to inhibit platelet aggregation and inflammatory responses²⁴³.

Parasites may also avoid or manipulate the host immune response by interacting with immunoglobulins (antibodies). In *S. mansoni* and *S. japonicum*, respectively, tetraspanins Sm23 and Sj23 are surface-exposed proteins that bind the Fc domain of human IgG antibodies and may help the parasites to evade host immune recognition or complement activation^{64,244}. A similar ‘non-immune’ (Fab-domain independent) binding of host IgG has been seen in bacterial pathogens such as streptococci²⁴⁵. Based on analysis of the *S. mansoni* genome, tetraspanin expansions have been described²⁴⁶ and in cestodes, we saw a tetraspanin family containing human CD63 that had expanded (a subfamily of all tetraspanins; family 306209; 9-12 copies per *Echinococcus* species, 6-16 in *Hymenolepis*, 1-2 in *Schistosoma*; **Supplementary Fig. 9b** and **Supplementary Table 9a**). The many tetraspanin subfamilies of animals were placed by Compara into separate Compara families, due to their divergence, but the Compara family that was expanded in cestodes also contained the Fc-binding *S. mansoni* Sm23 and *S. japonicum* Sj23 proteins. We also saw an expansion of a second, divergent tetraspanin Compara family in trematodes (family 730503; ~3-5 copies per *Schistosoma* species; 11 *Clonorchis sinensis*; 3 *Fasciola hepatica*; **Supplementary Fig. 9c** and **Supplementary Table 9a**).

Since lipids act as effectors and second messengers in the vertebrate immune system²⁴⁷, interactions with host fatty acids and retinoids may have immunomodulatory effects. In strongylids (clades Va, Vb, Vc) we saw an expansion of the fatty acid- and retinoid-binding protein (FAR) family (family 179875; 75% predicted secreted; 14-18 copies per *Ancylostoma* species; 8-10 *Haemonchus*, compared to 6 *C. elegans*, 2 *Globodera pallida*; **Supplementary Fig. 9d** and **Supplementary Table 9a**) that has previously been studied in plant parasites⁶⁶. FAR proteins are secreted by adult *Ancylostoma*²⁴⁸ as well as filaria²⁴⁹, and FAR genes are highly transcribed in adult *Haemonchus* and *Necator*^{67,250}. Their role is unclear, but this Compara family also included the *Globodera* gene *gp-far-1* (also called *sec-2*²⁵¹), encoding a surface protein that binds two lipids (linolenic and linoleic acids) that are precursors of plant defence compounds and the jasmonic acid

defence signalling pathway⁶⁶. Thus, secreted strongylid members of this family may play a role in eluding vertebrate host defences; although other possible roles include acquisition of lipids from the hosts, or delivering or sequestering signalling lipids and so modifying host tissues in which the adult parasites live⁶⁷.

Large numbers of glycosyl hydrolases related to degradation of glycosaminoglycans were seen in platyhelminths and nematodes, including GH20, GH27, GH56, GH84 and GH85 class enzymes (**Supplementary Fig. 30; Supplementary Table 25**). These enzymes degrade hyaluronic acid, chondroitin and N-acetyl glucosamine that give structural integrity to animal tissues²⁵²⁻²⁵⁴. A chondroitin hydrolase (GH56; EC 3.2.1.35) family was expanded in nematode clade Vc (Ancylostomatidae + Strongylidae) and with relatively high counts in other gastrointestinal parasites (family 117102; 48% predicted secreted; 15-21 copies in *Ancylostoma* species, 8 *Necator americanus*, 6 *Strongyloides stercoralis*; **Supplementary Fig. 9e; Supplementary Fig. 30 and Supplementary Table 9a**). This family included *C. elegans chhy-1*, a chondroitin hydrolase that is orthologous to human *HYAL1* and other hyaluronidases, and which cleaves chondroitin and chondroitin sulphates²⁵⁵. These enzymes may play a role in breakdown of host intestinal epithelium by the parasites, or for larval migration through host connective tissue. We also observed a small expansion of an uncharacterised GH5-class glycosyl hydrolase in schistosomatids (family 627564; 56% predicted secreted; 7 *S. mansoni*; **Supplementary Fig. 9f; Supplementary Fig. 30 and Supplementary Table 9a**). The glycosyl hydrolase GH5 class includes a variety of specificities; some (e.g. β -glucosidase) have previously been reported in the excretory-secretory products of trematodes (e.g. *Fasciola*), and may be involved in degrading supramucosal gels found on host epithelial surfaces²⁵⁶. This Compara family (627564) included a *S. haematobium* gene (*A_02570*) and a *S. mansoni* gene (*Smp_187070*) whose expression is highly enriched in eggs^{8,70}. An intriguing possibility is that this family is used by schistosomatid eggs to traverse the bladder or intestinal walls, depending on schistosome species, in order to be passed into the environment. Alternatively, these transcripts may be present in eggs, but translated at a subsequent stage of the life cycle, for example for miracidial remodelling or penetrating the snail.

Some helminth parasites such as *Trichinella* and *Trichuris* can live as intracellular parasites within host cells^{257,258}. In clade I, which includes these species, there has been an expansion of a family with the PAN/Apple domain, involved in carbohydrate/protein binding (family 211141; 64% predicted secreted; 5 *T. spiralis*, 4 *T. muris*, 6 *Romanomermis*; 2 *Soboliphyme*, compared to 2 *C. elegans*, 2 *B. malayi*, 2 *S. ratti*, 3 *A. suum*; **Supplementary Fig. 9g and Supplementary Table 9a**). Genes with the PAN/Apple domain are upregulated in the *T. muris* parasite larva compared to the adult¹⁷. One possibility is that these proteins play a role in host cell attachment and invasion in *Trichinella* and *Trichuris*, since in the intracellular protozoan parasite *Toxoplasma gondii*, a PAN/Apple domain protein binds host galactose-containing ligands and may enable *T. gondii* to attach to vertebrate host cells, facilitating host cell invasion⁷¹. Alternatively, the family may modulate galectin-mediated activation of the host immune system, as has been proposed in *T. gondii*²⁵⁹.

Gene family expansions related to the helminth surface coat and protective glycocalyx.

The surface coats of helminths (tegument or cuticle) offers protection from harsh environmental conditions and microbial pathogens, as well as from the host immune system. For example, larvae of *Echinococcus* tapeworms are protected by a thick layer comprising mucins that bear galactose-rich carbohydrates²⁶⁰, while the surface of *Taenia* oncospheres (larvae) are covered with carbohydrate chains rich in D-mannose, D-glucose, D-galactose and N-acetyl-D-galactosamine²⁶¹. A total of 9,947 glycosyltransferases which belong to a wide range of GT classes (49 classes) were identified in our analysis (**Supplementary Table 25; Supplementary Fig. 30**), suggesting these species have an ability to synthesise a wide range of glycans.

We observed a large expansion of a galactosyltransferase (GT31) family in cestodes, previously reported in *Spirometra*¹⁴ and cyclophyllidean tapeworms¹³ (family 100585; 36% predicted secreted; 20-30 copies per species; **Supplementary Fig. 10a**) and another GT31 family expansion in nematode clade Vc (family 540548; 7-20 copies in *Ancylostoma* species; **Supplementary Fig. 10b** and **Supplementary Table 9a**). In nematode clade IVa, and part of clade IVb (*Bursaphelenchus*, *Panagrellus*), a GT49 glycosyltransferase family was expanded (family 140262; 36% predicted secreted; 11-18 copies per species; **Supplementary Fig. 10c** and **Supplementary Table 9a**). The outer layer of the cuticle surface of many species of nematodes is covered by a carbohydrate-rich glycocalyx or surface coat⁶⁸. There are several plausible explanations for the expansion of the glycosyltransferase families, including: modification of the glycocalyx to provide resistance to microbial pathogens of nematodes that bind to glycosylated proteins of the glycocalyx, such as mucins²¹⁷; modification of the glycocalyx as a form of antigenic variation to help prevent immune recognition by hosts (for parasitic nematodes the glycocalyx and ES antigens present the main immunogenic challenge to the host⁶⁸); production of host-like glycans that modulate the host immune response by interacting with host proteins, such as lectins⁶⁹; or glycosylation of secreted adhesion molecules, such as those of *Strongyloides* species²⁶², that enable parasites to adhere to host epithelia.

Nematodes are protected by a tough surface cuticle, while nematode eggs are partly composed of the tough and resilient polysaccharide chitin. Glycosyl hydrolases related to chitin metabolism (classes GH18 and GH19), as well as chitin-binding enzymes (CBM14) were enriched in nematode species and only a few were found in trematodes and cestodes, probably reflecting the fact that chitin is one of the structural polysaccharides of nematode bodies and eggs (**Supplementary Table 25; Supplementary Fig. 30**). For example, a chitin-binding (CB14) Compara family was expanded in *Ascaris* and *Toxocara* (family 785119; 62% predicted secreted; 8-10 copies in *Ascaris* species, 13 in *Toxocara*; **Supplementary Fig. 10d** and **Supplementary Table 9a**), perhaps related to the particularly thick chitinous layer of *Ascaris* egg shells²⁶³, which may underlie their ability to persist in the soil up to ten years²⁶⁴. Analysis of synapomorphic gene families (**Supplementary Note 2.3**) recovered one chitin binding family restricted to clade III and four additional families specific to clade IIIa (Oxyuridomorpha). Trichocephalida (which includes *Trichuris* and *Trichinella*) also exhibited a synapomorphic chitin binding family, in addition to a GH18 family.

2.5 SCP/TAPS protein family

Authors: Huei-Mien Ke, Tzu-Hao Kuo, Tracy J. Lee and Isheng Jason Tsai

The SCP/TAPS (Sperm-Coating Protein/Tpx/Antigen 5/Pathogenesis-related protein-1) protein family (defined by presence of a Pfam CAP domain PF00188, or InterPro IPR014044) is ubiquitously found in both prokaryotes and eukaryotes⁷² and bears many alternative names: including cysteine-rich secretory proteins (CRISPs) and *S. mansoni* Venom Allergen-like proteins (SmVALs). Their functions are wide-ranging but are poorly understood^{72,265}. In parasitic worms, SCP/TAPS have been found as major components of excretory-secretory products, are expressed at key points in parasite life-cycles^{18,266-268} and there is mounting evidence of their role in modulating host immune responses^{269,270}. Examples of SCP/TAPS proteins are being developed as vaccine candidates²⁷¹⁻²⁷⁴.

There was a total of 3,915 putative SCP/TAPS genes present in the data set of 81 helminths (**Supplementary Table 10**). After filtering by sequence length (**Methods 10**), 3,167 SCP/TAPS sequences from the 81 helminths (3,121 helminth sequences) and two outgroups (15 from human, 31 from *D. melanogaster*) were clustered by USEARCH¹⁵⁰ using a threshold of $\geq 70\%$ protein sequence similarity. Only 1,456 (46.0%) SCP/TAPS sequences could be placed into 498 clusters, suggesting high divergence of these genes. A maximum likelihood phylogeny was computed based on an alignment of the other 1,711 singleton sequences and the consensus sequences of the 498 clusters. The phylogeny suggested that SCP/TAPS comprise two major subfamilies as previously proposed²⁶⁵ ('Group 1' and 'Group 2' in **Fig. 3, Supplementary Fig. 11**).

SCP/TAPS expansions in *N. americanus* (120), *Strongyloides* and *Parastrongyloides* (40-179 copies in five species), *Ancylostoma ceylanicum* (269), and *S. mansoni* (33) have previously been reported^{18,27,268,275}. Here, cross-phyla comparisons again revealed several independent expansions, particularly in parasitic clade V (18-381 copies in 17 species) and parasitic clade IVa intestinal nematodes (39-166 copies in five species; **Supplementary Fig. 11**), with the highest, 381 copies, identified in *Ancylostoma caninum*. This number might be somewhat inflated by unresolved heterozygosity in the *A. caninum* genome assembly (CEGMA (partial) gene count 1.49; **Supplementary Table 2a**). *Dracunculus medinensis* was unusual in being the only member of clade III nematodes to have undergone expansion, with 66 copies.

Of the 3,121 length-filtered SCP/TAPS genes in the 81 helminths, 79.4% were annotated with one single CAP domain, and no other domains. 10.4% (326) of genes were annotated with two SCP domains (**Supplementary Fig. 31**). The other 10.2% had other combinations (e.g. a single CAP domain plus some other domain(s)). The two-SCP/TAPS-domain genes were most commonly found in species with expansions. The two-SCP/TAPS-domain genes did not share a common ancestor within or between different species and were found proximal to single-SCP/TAPS-domain genes, suggesting independent fusions after gene duplication within different species.

2.6 GO annotation enrichment

Authors: Bruce A. Rosa, Avril Coghlan and James Cotton

GO terms were transferred to helminth proteins from human, zebrafish, *C. elegans* and *D. melanogaster* orthologues (**Methods 6**). Enrichment of GO terms or InterPro/Pfam domains in species groups of interest (e.g. clade III, Ascaridomorpha, etc.; see 'Analysis group' and 'Broader analysis group' in **Supplementary Table 4**) was tested, compared to a background of all species (**Methods 12**). This enabled us to identify some domains that are clade-specific or nearly so. For example, we identified only one InterPro Domain (IPR000791, acetate transporter GPR1/FUN34/SatP) which is present in all seven clade I nematode species, and none of the other species ($P=0.005$; **Supplementary Table 24**). This domain is found in 3 *Romanomermis* proteins, 4 from *Soboliphyme*, and 3 per *Trichinella* species and 3-4 per *Trichuris* species. Most of these proteins belong to one Compara family (family 972857), which only has clade I genes. This family was also identified in our analysis of synapomorphies (clade-specific families; **Supplementary Note 2.3**). Proteins containing the IPR000791 domain have roles in acetate/succinate permeability/transport in fungi and *E. coli*^{276,277}. In *Chlamydomonas* proteins containing this domain are upregulated in response to a boost of acetate in the medium²⁷⁸. Therefore, these proteins may be used by clade I species for acetate or succinate uptake, or efflux. For example, malate dismutation (which produces succinate and/or acetate) has been observed in clade I species (e.g. *Trichuris vulpis*²⁷⁹), so these transporters may possibly be used to remove acetate/succinate from malate dismutation that has built up. On the other hand, they may have a different ligand; some members of the GPR1/FUN34/SatP family have been suggested to be ammonium/H⁺ antiporters in yeast²⁸⁰. The IPR000791-domain is not usually found in animals, so it seems likely that the clade I IPR000791 genes were gained by horizontal transfer from fungi or bacteria. Indeed, taking one of the *T. muris* genes in this family (*TMUE_s0096004700*), we found its top non-clade-I BLASTP hits in the NCBI protein database were to a GPR1/FUN34/SatP family protein in the bacterium *Azoarcus* sp. KH32C (accession WP_015437673.1; 46% identity, *E*-value 4e-50), a β -proteobacterium. The second top non-clade-I BLASTP hit was also to a β -proteobacterium (*Rhodoferrax*, 45% id.), but other of the top five non-clade-I BLAST hits were to γ -proteobacteria (*Amphritea*, 45% id.) and α -proteobacteria (*Pararhodospirillum*, 45% id.) (**Supplementary Fig. 23b**). Interestingly, the *Chlamydomonas* IPR000791-domain genes seem to have been independently gained from bacteria²⁷⁸. Phylogenetic analysis of the nematode proteins (**Supplementary Fig. 23b**) confirmed that the clade I proteins form a monophyletic group distinct from either the *Chlamydomonas* hits or the proteobacterial hits. This analysis thus failed to confirm the origin of these nematode proteins from any specific bacterial group, possibly because the nematode proteins are too divergent from any prokaryotic sequences.

There are 4 *T. muris* genes with the IPR000791 InterPro domain (*TMUE_s0073001000*, *TMUE_s0003015100*, *TMUE_s0071005300*, *TMUE_s0096004700*). These are all on long scaffolds with many genes (37, 175, 71, 51, 71, 175 genes, respectively). *T. muris* *TMUE_s0073001000* is flanked by neighbouring/nearby genes that belong to Compara families containing genes from many nematodes including *C. elegans* (*C. elegans daf-16* and *daj-27*); as is

TMUE_s0003015100 (*C. elegans* C06B8.7 and *sptl-2*), TMUE_s0071005300 (*C. elegans* C50F7.6 and F29G6.1), and TMUE_s0096004700 (*C. elegans* *prp-31* and C05D10.4). It is therefore unlikely that these huge scaffolds are due to contamination of the assembly with endosymbiont DNA.

The GTP cyclohydrolase I feedback regulatory protein (IPR009112) was among the InterPro domains most significantly expanded in clade V species ($P=1\times 10^{-6}$), and was identified only in all clade V species, 7 outgroup species and one clade I nematode (*R. culicivora*) (**Supplementary Table 24**). GTP cyclohydrolase I feedback regulatory protein (GFRP) regulates GTP cyclohydrolase I and thereby synthesis of the cofactor tetrahydrobiopterin, which is required for hydroxylation of aromatic amino acids²⁸¹ and in biosynthesis of serotonin, and has been shown to be required for cuticle integrity and biogenic amine synthesis in *C. elegans*²⁸². Genes with this domain include *C. elegans* *grfp-1*, which belongs to a Compara family that only has clade V and outgroup species (family 1049829), containing human GTP cyclohydrolase I feedback regulator (*GCHFR*). *Drosophila* appears to have lost the *GFRP* gene, instead using a different mechanism to regulate GTP cyclohydrolase I²⁸³. Thus, it appears that most nematodes (except clade V) and platyhelminths have also lost the *GFRP* gene, so they may also have a different mechanism for regulating tetrahydrobiopterin synthesis.

3. Proteins historically targeted for drug development

3.1 Proteases and protease inhibitors

Authors: Avril Coghlan, Nancy Holroyd and Neil Rawlings

High expression of proteases and protease inhibitors in transcriptomic and proteomic data from parasitic stages of nematodes has been observed (e.g. astacins in *Strongyloides*¹⁸), as has expansion of some protease and protease inhibitor (PI) families (e.g. chymotrypsin A-like serine proteases in *Trichuris*¹⁷). We investigated helminth proteases in several ways: through the MEROPS database¹⁵², and during analysis of expanded gene families (*Gene family expansions: Supplementary Note 2.4*), novel domain combinations (*Novel domain combinations: Supplementary Note 3.2*) and clade-specific families (*Synapomorphies: Supplementary Note 2.3*).

Proteases and PIs were annotated using BLAST searches against MEROPS (**Methods 11, Supplementary Table 11**). Twenty-three protease families were found in every platyhelminth and nematode genome. Platyhelminths generally contained fewer genes that encode protease homologues than nematodes (medians 281 and 452; **Supplementary Table 11a, Fig. 4**). Looking at the higher quality tier 1 genomes (those with high-quality assemblies; **Methods 2**), on average platyhelminths had 306 protease genes per species (5-95% quantile range 228 to 429), compared to 485 in nematodes (283 to 766). This agreed with our general finding of a lower number of enzyme annotations (EC numbers) per platyhelminth species compared to nematodes

(**Supplementary Note 4.1**). Clade V species had more proteases on average (629 in tier 1 species) than any other nematode clade.

Astacins

The MEROPS protease family/subfamily with the most paralogues in any one genome, and the only family where there were more than a hundred paralogues predicted to be active, was M12 (astacin metalloendopeptidases). Astacins have previously been found to be expanded in *Strongyloides* and *Parastrongyloides*¹⁸. The ability of a recombinant astacin-like metallopeptidase from *Ancylostoma* to digest collagen⁷⁴, and the presence of astacins in the excretory secretory products of *Strongyloides* infectious-stage larvae²⁸⁴ suggest that astacins are involved in parasite migration through host tissue. Clade IVa and clade V species had the most paralogues: for example, *Parastrongyloides trichosuri* (clade IVa) had 331 and *Ancylostoma duodenale* (clade V) had 163.

In our analysis of gene family expansions (**Supplementary Note 2.4**), we identified six Compara families of astacins that had expanded, four of which were in clade IVa, one in clade Vb (strongylid-lungworm) and one in Vc (hookworms) (**Supplementary Table 9a; Supplementary Fig. 14a-g**). The family that was expanded in Vc (family 163808; 22 *O. dentatum*, 20 *N. americanus*, 13 *A. duodenale*, 12 *H. bakeri*, 11 *S. vulgaris*, 9 *A. caninum*, 8 *A. ceylanicum*) included *A. caninum* astacin *mtp-2*, which is secreted in adult ES products and is hypothesised to function in digestion of the host tissue plug lodged in the buccal capsule of the adult parasite²⁸⁵. *Ancylostoma caninum* astacin *mtp-1*, secreted by iL3s and involved in tissue migration⁷⁴, belonged to a Compara family that has independently expanded in clades V and IVa, and was not detected by our analysis of gene family expansions (family 3: 11 *T. circumcincta*, 76 *S. papillosus*, 72 *A. caninum*, 53 *H. bakeri*, 41 *H. contortus*, 35 *P. pacificus*, 31 *S. ratti*). *Strongyloides stercoralis* ‘strongylastacin’, which is secreted by iL3s and hypothesised to be involved by larvae of *S. stercoralis* to penetrate human skin²⁸⁶, belonged to family 132363, one of the four families identified as expanded in clade IVa.

In contrast to the nematodes, most platyhelminths had far fewer astacin homologues: *Fasciola hepatica* had only three (**Fig. 4**). On average, platyhelminths had just 7 astacin homologues, compared to an average of 62 across nematodes (tier 1 species). This was consistent with the hypothesis that nematodes use astacins for skin invasion and migration through host tissue but that platyhelminths use different proteases for these purposes, for example serine or cysteine proteases by schistosomes²⁸⁷.

Cathepsin proteases

Cathepsin proteases, members of the C1 (papain) family, have previously been implicated in host-parasite interaction phenomena such as tissue invasion⁷³, feeding²⁸⁸, and immune evasion²⁸⁹. The C1 family was one of 23 MEROPS protease families present in all helminth species.

In our analysis of expanded gene families (**Supplementary Note 2.4**), we identified expansions of two Compara families with the Pfam ‘Papain family cysteine protease’ (PF00112) domain, both

expanded in clades Va (strongylid-other) and Vc (hookworms): family 38173 (41 *H. contortus*, 40 *O. dentatum*, 33 *T. circumcincta*, 26 *Strongylus vulgaris*, 20 *A. caninum*; **Supplementary Fig. 14h**) and family 80624 (56 *O. dentatum*, 36 *A. caninum*, 25 *S. vulgaris*, 24 *H. contortus*, 19 *T. circumcincta*; **Supplementary Fig. 14i**). The former family (38173) included human cathepsin B (CTSB), and had members in most helminths, including several known cathepsin B genes from *C. elegans* (e.g. *cpr-6*, *cpr-8*), *H. contortus* (e.g. AC-2²⁹⁰ and AC-4²⁹¹), *Fasciola hepatica* (e.g. *FhcatB1*²⁸⁹), and *S. mansoni* (e.g. *Sm31*²⁹²). The other Compara family (80624) was found primarily in clade V, and did not include a human member, but did have several cathepsin B genes from *C. elegans* (*cpr-1*, *cpr-2*, *cpr-3*, *cpr-4*, *cpr-5*), *H. contortus* (e.g. *gcp-7*²⁹³ and *hmcp-2*²⁹⁴) and *Ancylostoma caninum* (e.g. *CP-2*²⁹⁵). There were few genes from flatworms in Compara family 80624 (e.g. 4 *F. hepatica*, 1 *S. japonicum*, no *S. mansoni*) but in family 38173 there were several (e.g. 11 *F. hepatica*, 10 *S. japonicum*, 9 *T. regenti*, 9 *S. mansoni*). This confirmed previous findings of a large number of cathepsin B genes in *H. contortus* (63 genes³⁰), and marked expansions of cathepsin B genes in schistosomatids and fasciolids¹². Note that while our two Compara families both had a mixture of *C. elegans* and *Haemonchus* cathepsin B genes, a phylogenetic analysis by Laing *et al*³⁰ found that *Haemonchus* genes formed separate clades from those of *C. elegans*.

A striking association between these two Compara families and the haematophagous nature of some adult helminths was evident. The nematode species that were most abundant in the families are blood-feeding (*Haemonchus* in clade Va; *Ancylostoma*, *Oesophagostomum*, and *Strongylus* in clade Vc), as were the most abundant platyhelminth species (the flukes *Fasciola*, *Schistosoma* and *Trichobilharzia*; **Supplementary Table 12**). Many of the proteases in these families are most highly expressed in adult worms (e.g. *H. contortus* AC-2²⁹⁰, AC-4²⁹¹, *gcp-7*²⁹³, *hmcp-2*²⁹⁴ and *A. caninum* CP-2²⁹⁵), and larval stages within the host (e.g. *H. contortus* *cpr-6* homologues²⁹⁶, *O. dentatum* cathepsin Bs²⁸) and are hypothesised to play a role in blood feeding²⁹⁷. For example, *A. caninum* CP-2 is part of a haemoglobin degradation pathway⁷⁵, and, AC-4 is the main cysteine protease in fibrinogen-degrading extracts purified from adult *H. contortus*²⁹¹. Likewise, *S. mansoni* *Sm31* is expressed in adult worms²⁹² and involved in degrading blood²⁹⁸; cathepsin Bs are part of a blood degradation cascade in *S. mansoni*⁷⁶.

In addition, it is likely that some of the cathepsin B proteases expressed in specific developmental stages may be involved in functions unrelated to blood feeding. *F. hepatica* *FhcatB1* is thought to play a role in the digestive tract of newly excysted juvenile parasites²⁹⁹, while other *F. hepatica* cathepsin B proteases may be involved in metacercariae excystment, invasion of the host^{78,300}, penetration of the host intestinal wall³⁰¹, and immune evasion by inducing degradation of immunoglobulin heavy chain²⁸⁹. In nematodes cathepsin Bs may also play other roles besides blood feeding; for example, a cathepsin B might be involved in larval development, and possibly emergence of the nematode *Parelaphostrongylus* from its intermediate snail host⁷⁷.

Trypsin proteases, prolyl oligopeptidases, and aminopeptidases

In addition to the expansions of astacins and cathepsins described above, our Compara analysis (**Supplementary Note 2.4**) also detected the previously described expansions of prolyl oligopeptidases in clade IVa¹⁸, trypsin proteases in clade I¹⁷, and aminopeptidases in cestodes¹⁴.

Trypsin is secreted from the intestinal infective larvae of *Trichinella spiralis*³⁰². *Trichuris muris* adults strongly express trypsin domain proteins in larvae, and in the anterior region of adults, and most are probably secreted¹⁷. Trap *et al*³⁰³ suggested that *Trichinella spiralis* trypsin domain genes may be involved in proteolysis of cuticle proteins during larval moulting, and/or digestion of host tissue for nutrition. The trypsin proteases are serine proteases (MEROPS family S1) that showed an expected large expansion in clade I, with the exception of *Romanomermis* that unlike the other clade I species sampled, is a parasite of insects rather than vertebrates (**Fig. 4; Supplementary Table 12**).

A proteomic analysis found an abundant prolyl oligopeptidase that is specific to the excretory and secretory (ES) products of female *S. ratti* worms²⁸⁴. Prolyl oligopeptidases (MEROPS family S9) showed a large expansion in clade IVa (**Fig. 4**).

A previously observed expansion of MEROPS family M17 (amino-terminal leucyl aminopeptidases) in the *Spirometra* genome¹⁴ was also observed in the cestodes *Dibothriocephalus* and *Mesocostoides* (**Fig. 4**). The expansion could relate to the complex life cycles of these cestodes that include the stages for colonising secondary intermediate hosts.

Protease inhibitors

Only four MEROPS inhibitor families were found in all nematode and platyhelminth proteomes (**Supplementary Table 11b**): I1 (Kazal), I2 (Kunitz-BPTI; see below), I4 (serpin) and I93 (sizzled), which inhibit a range of proteases (e.g. serine proteases, cysteine proteases, metalloproteases). There were several MEROPS inhibitor families commonly present in nematodes but rare or missing in platyhelminths, for example: I17 (see below), I12, I19, I31, I33, I35 and I83. In contrast, there were no MEROPS protease inhibitor families prevalent in platyhelminths that were rare or absent in nematodes (**Supplementary Table 11b; Fig. 4**).

Trypsin, trypsin-like, and chymotrypsin/elastase protease inhibitors

Different families of trypsin inhibitors can be used to protect helminths from degradation by host proteases, to facilitate feeding and to manipulate the host response to the parasite⁷⁹.

I2 inhibitors. The most abundant protease inhibitor across all parasitic nematodes and platyhelminths was I2 (Kunitz-BPTI), a trypsin inhibitor present in all helminth proteomes (**Fig. 4**). I2 inhibitors inhibit proteases of the S1 serine protease family, which includes trypsin and chymotrypsin. Nematodes had an average of 102 I2 members per species, whilst platyhelminths had 38.

Kunitz-type (I2) inhibitors are expanded in cestodes and especially *Spirometra*¹⁴. In agreement with this, we found an expanded Compara family of Kunitz-type inhibitors in cestodes (family 576282; 10 per *Echinococcus* species; 6-9 per *Taenia* species; 9 *Spirometra*; **Supplementary Fig. 14n; Supplementary Table 9a; Supplementary Note 2.4**). Although MEROPS family I2 was found in all flatworms, it was most abundant in cestodes (**Fig. 4**).

A putative diverged trypsin inhibitor Comparsa family (family 616925), distantly related to Kunitz-BPTI inhibitors, was present and expanded only in clade Va and Vc species, and absent from lungworms (14 *H. bakeri*, 8 *H. contortus*, 8 *A. caninum*, 7 *S. vulgaris*, 6 *T. circumcincta*; **Supplementary Fig. 14j**; **Supplementary Table 9a**). The family had low sequence similarity to Kunitz/BPTI; 11 genes in the family had weak BLASTP hits in the NCBI protein database to an *Ancylostoma* protein annotated as 'Kunitz/Bovine pancreatic trypsin inhibitor domain protein' (accession EPB75542.1; e.g. BLASTP hit from ANCCAN_24592 had 28% identity, *E*-value 7e-13). No known MEROPS PI family was identified for this family, suggesting these could be a novel protease inhibitor specific to clade V parasites (other than lungworms).

I8 inhibitors. Many trypsin-like inhibitor genes are upregulated in the parasitic females of *Strongyloides ratti*¹⁸. Consistent with this, in clade IVa we identified an expanded Comparsa family of trypsin inhibitor-like genes (family 426447; 17 *P. trichosuri*, 9-14 per *Strongyloides* species; **Supplementary Table 9a**; **Supplementary Fig. 14l**). Genes in this Comparsa family (e.g. SRAE_1000260200) were classified in the MEROPS 'chymotrypsin/elastase inhibitor' family (I8), which inhibits trypsins, chymotrypsins and elastases and was abundant in clades IVa and Vc, and to a lesser extent in clade IIIb (**Fig. 4**).

In clade IIIb we also observed an expanded chymotrypsin/elastase (I8) inhibitor Comparsa family (family 754923; 8 per *Ascaris* species, 3 *Toxocara*). Chymotrypsin/elastase inhibitors from *Ascaris* are thought to protect it from host proteases⁸¹ (**Supplementary Fig. 14m**).

Other protease inhibitors

I17 inhibitors. The previously reported expansion of the protease inhibitor family I17 (WAP-type inhibitors such as antileukoprotease, inhibitors of trypsin-like serine endopeptidases) in clade I¹⁷ was even more striking in this larger data set (**Fig. 4**). I17 was commonly present in nematodes but rare or missing in platyhelminths.

a2M inhibitors. Among our Comparsa family expansions (**Supplementary Note 2.4**), one of the most prominent expansions observed was a protease inhibitor (MEROPS family I39), α -2-macroglobulin (a2M) homologue (Comparsa family 358015). MEROPS family I39 members interact with a wide variety of endopeptidases such as aspartic and serine proteases. This Comparsa family was present in all platyhelminth species (but not nematodes), and a significant expansion was observed in cestodes, particularly pseudophyllidea (15 *S. erinaceieuropaei*, 15 *D. latus*, 6-8 per *Taenia* species, 3 per *Echinococcus* species, 2-13 per *Hymenolepis* species; **Supplementary Table 9a**, **Supplementary Fig. 14k**). The only *C. elegans* gene with the Pfam α -2-macroglobulin domain (PF00207), *tep-1*, was present in a different Comparsa family (172243), with many nematode members with 1-6 copies per species. In agreement with this, the MEROPS family I39 (α -2-macroglobulin, or a2M) was present in most helminths, including nematodes, but most abundant in cestodes (**Fig. 4**). In flatworms, some α -2-macroglobulin genes were found to be highly expressed in *O. viverrini* stages within the biliary tract of the host³⁰⁴. Vertebrate α -2-macroglobulin is known to inhibit coagulation by inhibiting thrombin³⁰⁵, and in *Opisthorchis*, reduction of clotting at attachment or feeding sites³⁰⁶ may be aided by this gene. α -2-macroglobulin

has also been shown to bind to several important cytokines and hormones⁸⁰. This expanded family could therefore potentially be important in immunomodulation, by reducing the activity of defensive signalling proteins from the host.

3.2 Novel domain combinations, including those related to proteases

Authors: Avril Coghlan, Adam Reid and Myriam Shafie

We found 14,596 combinations of 4216 different Pfam domains that were present in nematodes or platyhelminths, but absent from other phyla in UniProt. Of these, 131 Pfam domain combinations were specific to nematodes (**Supplementary Table 14a**), and 50 specific to platyhelminths (**Supplementary Table 14b**).

In line with our findings more generally across these species (see *Proteases and Proteases Inhibitors*, **Supplementary Note 3.1**) we identified novel phylum-specific domain combinations involving several protease domains, in both phyla (**Supplementary Table 14**). Nematode-specific protease-related domain combinations included the ShK domain-like and Carboxypeptidase activation peptide domain, EB module domain and Kunitz/Bovine pancreatic trypsin inhibitor domain, and the Kunitz/Bovine pancreatic trypsin inhibitor domain and EGF-like domain. These were all found in both parasitic and free-living nematodes, so are probably not related to parasitism. For example, the ShK-like and Carboxypeptidase activation peptide domain combination was found in several *C. elegans* genes, including *suro-1*, which regulates cuticle formation and body morphogenesis³⁰⁷.

We found four novel protease-related domain combinations specific to platyhelminths: the Ku70/Ku80 N-terminal alpha and beta domain with Prolyl oligopeptidase domain; Ankyrin repeats with the Caspase domain; Aminopeptidase I zinc metalloprotease (M18) domain with 'helix-turn-helix, Psq' domain; and Peptidase family M28 domain with Enhancer of rudimentary domain (**Supplementary Table 14**). These were not found in the free-living species *Schmidtea mediterranea*, so are possibly related to parasitism. For example, the combination of M18 metalloprotease and 'helix-turn-helix, Psq' (a domain involved in DNA binding³⁰⁸) was found in 15 genes belonging to the same Compara family (768925). This Compara family included *S. mansoni* *Smp_210600* and *Smp_210610*, adjacent gene predictions that each have one of these two domains, and were recently reannotated as a single gene *Smp_246800*. This gene was shown to have significantly higher expression in 3-hour post-infection schistosomula compared to cercariae¹⁰, suggesting a potential role in regulating gene expression shortly after invasion.

3.3 GPCRs

Authors: Tim Day, Nicolas J. Wheeler and Mostafa Zamanian

Known GPCRs from *C. elegans*, *B. malayi*, *O. volvulus*, *S. mansoni* and *S. mediterranea* (providing information from both parasitic and free-living flatworms/roundworms) were identified from literature mining and previous GO annotations and used as seeds for identification of GPCR

families in our Compara database (**Methods 11**). Additional GPCR families were identified from our analysis of synapomorphic gene families, and manual curation, resulting in a total of 230 GPCR families, which included 5,939 genes from our 33 'tier 1' (with high-quality assemblies; **Methods 2**) helminth species (**Supplementary Table 15a, b, d**). The number of GPCRs per species ranged from 58-252 (5-95% quantile) across the 33 tier 1 helminths.

A heatmap of these data uncovered a number of trends (**Supplementary Fig. 15**). First, the massive radiation of chemosensory GPCRs in *C. elegans* was unmatched by any other nematode species. Indeed, of the *C. elegans* complement of 1402 GPCRs, 87% were chemosensory GPCRs (**Supplementary Table 15b**). This did not seem to be due to better annotation of GPCRs in *C. elegans* compared to other species. By comparing to *C. briggsae*, it has been found that the high GPCR count of *C. elegans* is due to expansion, rather than loss from other lineages^{187,309,310}. After *C. elegans*, the next largest chemosensory fraction occurred in the clade IVa parasitic nematodes *Strongyloides papillosus* (48%) and *S. venezuelensis* (47%), then in *Parastrongyloides trichosuri* (43%) and the free-living *Rhabditophanes* sp. KR3021 (43%; **Supplementary Table 15b**). All examined parasitic nematodes possessed chemosensory receptors but, after normalising for the total numbers of genes, the greatest numerical representation was apparent in clade IVa parasites (**Supplementary Fig. 15**). In particular, there were several synapomorphic chemoreceptor families in this clade (**Supplementary Table 15a**), which likely reflects the extensive range of environments with which these nematodes interact, given their life cycles that involve full development of free-living adults as well as multiple parasitic stages. Some chemoreceptor families were almost completely conserved across all parasites in the phylum. These included homologues of *C. elegans* *daf-37*, which is known to mediate ascaroside signalling, consistent with the importance of this small molecule signalling pathway in the likely evolutionary transition from dauer larvae to infective larvae³¹¹.

Interestingly, no chemosensory GPCRs were readily identifiable in platyhelminth species, even though they are found in arthropods and vertebrates and diverged chemoreceptors are found in other phyla such as Molluscs⁸². Given that platyhelminths, like nematodes, must navigate diverse environmental cues, it is unlikely that these species do not engage in sophisticated forms of chemosensation. We hypothesise that either some subset of identified phylum-specific orphan receptors is involved in such processes, or that chemosensation in flatworm parasites is not primarily driven by GPCRs.

A difference was seen in the GPCR complements of filarial nematodes and non-filarial nematodes. Filarial nematodes are nested within clade IIIc (Spiruromorpha+other), but phylogenetically diverged from their non-filarial counterparts (**Fig. 1**). *O. volvulus*, *L. loa*, *B. malayi*, and *L. sigmodontis* showed an overall decrease in GPCR families and, on average, possessed about 41% fewer GPCRs than other members of clade III among the tier 1 species (medians 103 versus 145).

Within clade I, *Trichinella spiralis*, *Trichuris muris* and *Soboliphyme baturini* were similarly reduced in their GPCR complements (58, 49, 69 genes, respectively) compared to other non-filarial nematodes. In contrast, *Romanomermis culicivorax* appeared to contain a more standard GPCR

complement (174 genes). An earlier but similar cluster analysis of the *R. culicivorax* genome¹⁵, reported differences in gene content between the whole proteomes of clade I species. The authors surmised that *T. spiralis* is not representative of the rest of the clade, but the present GPCR analysis suggested that in terms of GPCR content, it is actually *R. culicivorax* that differs more.

Although many families are conserved in both phyla, we observed far fewer GPCRs in platyhelminths than in nematodes (medians 75 versus 182; 5-95% quantiles of 60-114 versus 58-263, for tier 1 species; two-sided Wilcoxon test, $P=0.005$). Platyhelminths also had fewer GPCR Compara families (57 versus 194 containing tier 1 species). However, there were 21 families that had members of both phyla. There were also a number of platyhelminth-specific groups, such as the PROFs (Platyhelminth-specific Rhodopsin-like Orphan Families). These receptors are Class A receptors and hypothesised to be peptide responsive but they do not show any significant homology to any annotated or deorphanised proteins³¹². Additionally, there were several other non-PROF GPCR families that were specific to flukes (trematodes), showing a clade-specific GPCR expansion (**Supplementary Fig. 15**).

3.4 Pentameric ligand-gated ion channels

Authors: Robin Beech and Mostafa Zamanian

The pentameric ligand-gated ion channels (pLGIC) mediate fast synaptic neuromuscular signalling and are the targets of many anthelmintic drugs³¹³. Five subunits combine to form either homo- or hetero-pentameric channels. The family can be divided broadly into the activating, cationic channels that lead to membrane depolarisation and inhibitory anionic channels that lead to hyperpolarisation and further divided based on their principal neurotransmitter, including acetylcholine, choline, biogenic amines, glutamate, GABA and protons among others.

The nematodes are particularly rich in pLGIC subunit genes, with the model nematode, *C. elegans*, encoding at least 29 cationic channel subunits and more than 35 anionic channel subunits³¹⁴. All parasitic nematode species had members of previously described nematode acetylcholine receptor classes (*deg-3*, *acr-16*, *unc-29*, *acr-8* and *unc-38*). We found both cationic and anionic receptor subunits had expanded in nematodes clades III, IV and V, following divergence from clade I (**Supplementary Fig. 16**; **Supplementary Table 16**). This suggested that neurotransmitter signalling is particularly important and diverse, with expanded, derived channels regulating behavioural characteristics specific to the nematodes such as chemotaxis, feeding, movement and reproduction.

In contrast, platyhelminths contained a more limited number of subunit genes that varied by class: *acr-16*-like and *acr-26*-like predominated in the flukes and *unc-63*-like in the tapeworms. The trematode, *S. mansoni*, encodes 27 different pLGIC subunits, four of which encode glutamate-gated anion channel subunits³¹⁵, and at least five are believed to encode acetylcholine-gated anion channels⁸⁴. A phylogenetic tree of all pLGIC channel subunits (**Supplementary Fig. 17**) placed the latter 5 in a distinct clade, most closely related to the nematode cationic channels *acr-26* and *acr-27*, that extended to all sequenced platyhelminths. Notably, an earlier, smaller phylogenetic

analysis suggested that the trematode acetylcholine-gated anion channels were related to another nicotinic acetylcholine receptor, *Clonorchis acr-16*⁸⁴, but this earlier analysis was based on fewer genes and did not include *acr-26/acr-27* representatives. *S. mansoni* acetylcholine-gated chloride channel genes in the previous analysis⁸⁴ belonged to Compara families whose members were placed in the *acr-26/acr-27* clade of our phylogenetic tree (families 921476, 123484). One of the previously studied genes, *Smp_142700*, was not included in our final tree due to poor alignment of the small Compara family to which it belongs, but preliminary phylogenetic analyses showed that it fell within the cationic *acr-26/acr-27* clade. The apparent switch of some acetylcholine receptor subunits from cationic to anionic is not unprecedented. For instance, the EXP-1 GABA receptor in *C. elegans* (that controls defaecation) switched from anionic to cationic, most likely in the ecdysozoan ancestor of the nematodes^{316,317}. A similar end result is achieved in the nematodes with a switch in ligand specificity of an anionic channel to bind acetylcholine³¹⁸.

A reversal in acetylcholine receptors from cationic to anionic in the platyhelminths should have led to dysregulation of a tightly coordinated system in any intermediate stage. It seems likely, therefore, that the role of acetylcholine in an ancestor was much reduced, with other signalling molecules, perhaps the biogenic amines³¹⁹, taking a more dominant role. The newly derived anionic acetylcholine receptor could then later acquire a more significant physiological role in the neuromusculature. Understanding how this could occur may be extremely illuminating for our understanding of how neuromuscular signalling has evolved.

3.5 ABC Transporters

Author: Robin Beech

The family of ATP dependent (ABC) transporters are multi-pass membrane proteins that actively transport a diverse range of substrates. Although they share structural similarities, their organisation falls into several distinct family categories³²⁰. The number of ABC transporter genes in nematodes was greater than vertebrates with 60 in *C. elegans*³²¹ and 40 in *S. mansoni* (**Supplementary Fig. 19; Supplementary Table 17**), compared to 49 in humans³²². For example, there were 14 functional p-glycoproteins in *C. elegans* (**Supplementary Table 17**) compared to four in humans⁸⁶. The families appeared dynamic, with gene gain and loss events changing the complement of orthologues found in different species, while overall the numbers remained relatively constant. Known as drug transporters from their link with resistance to chemotherapy in cancer treatment, the biological roles of p-glycoproteins specifically in helminths are not well understood. In *C. elegans* many are expressed in the pharynx, intestine and amphids and are believed to be involved in lipid transport and protection from environmental toxins, heavy metals and infection^{86,321,323,324}. Parasitic platyhelminths and nematodes spend a significant part of their life cycle within a host. The P-glycoproteins (pgp) may play a direct role in protection from the host immune system, which may explain their abundance in helminth genomes³²⁵.

There is particular interest that increased Pgp expression may lead to anthelmintic resistance. The evidence for this comes from the fact that inhibitors of Pgp activity reverse anthelmintic

resistance³²⁶ and that expression of nematode and trematode Pgp in mammalian tissue culture cells leads to anthelmintic resistance³²⁷. This is a complex issue since different Pgps may act in concert and it is difficult to ascribe resistance to individual Pgp genes. In addition, the family appeared dynamic with loss of some Pgp copies and expansion of others^{25,30} (**Supplementary Fig. 19**), which means information gained on function in a model such as *C. elegans* may not be translated directly to an understanding in related parasitic species.

The membrane phosphatidylserine is normally limited to the internal leaflet of surface membranes. The ABC transporter *ced-7* in *C. elegans* is involved in transport of phosphatidylserine to the outer membrane leaflet, where it signals cells for apoptotic body clearance³²⁸. Orthologues of this gene were found in the nematodes (**Supplementary Fig. 19**). Although distant homologues of *ced-7* were found in a few platyhelminth species (**Supplementary Fig. 19**), they belonged to different Compara families than the nematode *ced-7* genes, so were possibly not *ced-7* orthologues. The schistosome membrane contains phosphatidylserine³²⁹ so signalling the membrane for apoptotic clearance may be achieved through a different mechanism, or the flipping of phosphatidylserine may be achieved by other means.

3.6 The kinomes of nematodes and platyhelminths

Authors: John Martin and Bruce A. Rosa

The 81 nematode and platyhelminth species had 251 ± 63 (mean $\pm$ standard deviation) kinases per species (**Methods 11; Supplementary Table 23**). Tyrosine kinases (TK) (and tyrosine kinase-like, TKL, kinases), when combined, were the most abundant family (75 ± 21 per species; **Supplementary Fig. 18**), as previously reported for *C. elegans*²², but these were variable between nematode clades (54 ± 4 among clade I species, and 91 ± 26 among clade V species). The largest difference in abundance between nematodes and platyhelminth species was the Casein Kinase (CK1) family, as previously found in a comparison of *C. elegans* and *S. mansoni*¹⁸², with an average of 47 ± 18 genes among nematode species and only 13 ± 5 among platyhelminth species.

In platyhelminths, a few noticeable cestode-specific protein kinase expansions were observed in our analysis of Compara family expansions (**Supplementary Note 2.4**), including an A,G&C (AGC) family (id. 110854, which contains *C. elegans* *rsks-1*, implicated in longevity³³⁰; **Supplementary Fig. 18; Supplementary Table 9a**).

Two CK1 kinase families (8185, which includes *C. elegans* *ttbk-2* and *spe-6*, and 38771), and a TKL family (22454, which includes *C. elegans* *spe-8* and *frk-1*), were highly abundant and universally conserved across nematodes (with the exception of family 38771 in clade IIIa), and were also almost universally absent across all Platyhelminthes species (with the exception of a *Clonorchis* member in family 22454; **Supplementary Fig. 18, Supplementary Table 23**). The two CK1 families were both tau tubulin kinase-like families; in vertebrates tau tubulin kinases control the phosphorylation state of the microtubule-associated protein Tau, thereby regulating its binding to neuronal microtubule networks^{331,332}. However, *spe-6* and *spe-8* are involved in sperm activation

in *C. elegans*^{333,334}, and *frk-1* in cell division during development³³⁵, so these CK1 and TKL families may have different roles in nematodes.

4. Metabolic reconstructions of nematode and platyhelminth parasites

Authors: Rahul Tyagi, Swapna Lakshmipuram Seshadri, Avril Coghlan, James Cotton, and John Parkinson

4.1 Overall metabolic potential

Number of EC numbers and enzymes

Overall, 789 unique EC numbers were confidently ascribed to enzymes across the 81 genomes but there was a marked difference in the number that could be identified in nematodes versus platyhelminths (means 395 and 306 without hole-filling, respectively; t-test $P = 3.7 \times 10^{-8}$; **Supplementary Table 18a**; **Supplementary Fig. 20a**). Two hundred of these ECs were nematode-specific (i.e. not detected in any flatworm) and 93 were platyhelminth-specific (**Supplementary Fig. 32a, b**). The total EC-number count increased to 850, when pathway hole-filling was used to increase annotation sensitivity for the 33 tier 1 species (those with high-quality assemblies; **Methods 2**). Even when less reliable EC annotations from Compara families were included (**Supplementary Table 18b**), the most contiguous flatworm genomes had >60 fewer unique EC numbers than similarly highly contiguous nematode genomes (e.g. *S. mansoni* 437 and *E. multilocularis* 404 versus *S. rattii* 544, *T. muris* 498, *C. elegans* 607). This may be partly due to real biological differences (e.g. loss of enzymes from certain platyhelminth clades), but could also reflect the greater number of nematodes compared with platyhelminths in this study and the rich biochemical knowledge that already exists for *C. elegans* and *D. melanogaster* (another ecdysozoan), greatly facilitating enzyme prediction in nematodes compared with the more distantly related platyhelminths.

As expected, the total proteome size (total amino acids for all coding sequences) was significantly correlated both with the total number of enzymes and the number of unique EC numbers ($P < 10^{-5}$ and $P = 3 \times 10^{-5}$, respectively; **Supplementary Fig. 32c, d**). The extent to which KEGG helminth-relevant reference pathways (**Methods 13**) were covered in different taxonomic groups, defined as the fraction of ECs present (**Supplementary Table 18e-f** and **Supplementary Fig. 21a**), confirmed the apparently reduced metabolic potential of platyhelminths relative to nematodes: 74% of reference pathways had significantly lower coverage in platyhelminths. Amongst nematodes, many KEGG pathways (51%) appeared to have higher coverage among clade IVa parasites (**Supplementary Fig. 21a**). Four of the five clade IVa parasites in our analysis belong to the genus *Strongyloides*. This observation may partly reflect the biology of this genus (i.e. evolutionary loss of enzymes/pathways may be very rare in this genus), but also likely relates to their generally highly

contiguous assemblies. In fact, when less reliable EC numbers were considered, assigned by hole-filling or inferred from other Compara family members (**Supplementary Table 18a-b**), the number of unique EC numbers in the tier 1 clade IVa species (with a high-quality assembly; **Methods 2**) *S. ratti* approached that of the well-studied model *C. elegans* (544 versus 607). In contrast, filarial nematodes (representing clade IIIc in **Supplementary Fig. 21a**) had a dramatically reduced metabolic coverage compared to other nematodes, with 48% of KEGG reference pathways having significantly lower coverage, due to partial or complete loss of specific pathways (e.g. absence of the glycine cleavage system; see *Amino acid metabolism* below). Even *B. malayi* and *O. volvulus*, both tier 1 species, had >50 fewer unique EC numbers than *S. ratti* (492 and 483 versus 544, including those from hole-filling and Compara families; **Supplementary Table 18a-b**).

Variability in KEGG pathway coverage

In addition to their reduced KEGG pathway coverage, the pathways in platyhelminths showed significantly more variable coverage between species than seen between nematode species (**Supplementary Fig. 21b**, Wilcoxon test $P=0.004$). Amongst the platyhelminths, the flukes (trematodes) showed the highest variation in pathway coverage, whereas amongst the nematodes, clades IVb (which includes *Panagrellus* and the plant parasites) and IIIb had the most variation. This result did not change even when only a single species per genus was considered (data not shown), which suggests that the observed difference in variation is not entirely due to biases resulting from having multiple species from the same genus. Comparing the super-pathways using variation in pathway coverage across helminths, nucleotide metabolism pathways (only two pathways) and carbohydrate metabolism pathways showed the least variation in coverage while lipid and amino acid metabolism pathways showed much greater variation in coverage (**Supplementary Fig. 21c**). Some of these differences in pathway coverage variation may represent true biological differences (e.g. reflecting a wide range of lipid biochemistry), but others are likely to be because certain enzymes in these pathways or clades are more challenging to annotate. Indeed, the high variability in pathway coverage in platyhelminths probably reflected a greater variability in assembly quality (which will have affected EC-number annotation) compared to nematodes: the interquartile ranges of two measures of assembly quality, assembly contiguity (N50/scaffold-count) and CEGMA (partial) score, were both greater in flatworms than in nematodes (470 versus 249, and 6.9 versus 2, respectively, for tier 1 species; **Supplementary Table 2**).

Coverage of KEGG modules

Metabolic potential was also compared among helminths by taking pathway topology into account, by ascertaining the presence of a metabolic module based on annotation of all indispensable enzymatic steps, as previously described¹⁷⁴. Clustering based on 'complete' pathway modules identified in each species, in which all pathway steps had to be present (**Supplementary Table 18c-d**), confirmed the observations based on the topology-agnostic KEGG pathway coverage results described above, while providing some additional observations included in the pathway-specific results (below). Clustering species based on metabolic module presence produced a tree

generally consistent with the species tree (**Supplementary Fig. 32e**). Consistent with the common theme, platyhelminths tended to have reduced metabolic potential relative to nematodes in terms of the number of metabolic modules complete in the worms from the two phyla (mean 21.1 and 26.8 complete modules, respectively; t-test $P < 9e-4$; species-wise counts of complete modules shown in **Supplementary Fig. 32c, d**).

4.2 Lipid metabolism

Lipid metabolism pathways had the highest median coverage variability among helminths (**Supplementary Fig. 21c**) suggesting either that lipid metabolism genes are particularly hard to annotate (e.g. due to fast sequence divergence), or that there is variable sub-pathway representation due to different levels of dependence on lipids as food reserves. Lipids can be used as fuel through their degradation to fatty acids, and then to acetyl-CoA by the β -oxidation pathway. Lipids such as triglycerols are used as highly-concentrated energy reserves in nematodes from aerobic habitats, for example, during embryogenesis of *Ascaris lumbricoides* eggs²⁶³, and by infective larvae and free-living adults of *Strongyloides ratti*³³⁶.

β -oxidation

β -oxidation of fatty acids yields numerous ATP molecules when glucose intake is low. The KEGG β -oxidation module (M00087) was the only metabolic module that was complete in all nematodes, yet appeared incomplete in most platyhelminths, except for the liver fluke *C. sinensis* and the cestode *S. solidus*, based on the presence of at least one enzyme in the reaction cascade not being confirmed by our stringent annotation. This concurred with previous discovery of β -oxidation genes in the genomes of *C. sinensis* and its close relative *Opisthorchis*³⁰⁴, and was consistent with biochemical evidence for the β -oxidation enzymes in *S. solidus* (although they do not appear to oxidise exogenous palmitate)³³⁷. Furthermore, we identified putative β -oxidation genes in additional flatworms (**Supplementary Table 19a**) using EC annotations from a less stringent approach based on Compara families (**Supplementary Table 18b**). These included enzymes for all four steps of the β -oxidation pathway (acyl CoA dehydrogenase, enoyl CoA hydratase, L-3-hydroxyacyl CoA dehydrogenase, and β -ketothiolase) in the liver fluke *F. hepatica* (consistent with biochemical evidence³³⁸), free-living flatworm *S. mediterranea*, monogenean *P. xenopodis*, and the cestode *S. erinaceieuropaei*. Candidates for some of the four steps were also found in the fluke *E. caproni* and cestode *D. latus* (**Supplementary Table 19a**). The β -oxidation module also appeared nearly complete (just missing one step) in *Taenia solium* (**Supplementary Table 19a, Fig. 5a**). Consistent with this, a Compara family with members from *C. sinensis*, *F. hepatica*, *S. mediterranea*, *P. xenopodis*, *E. caproni* and *D. latus*, contained the *C. elegans* carnitine palmitoyl transferase (*cpt-1*) that transports long chain fatty acids into the mitochondrion for β -oxidation. Curiously, we did not identify β -oxidation genes, or *cpt-1* homologues, in several tier 1 flatworm species with highly contiguous genomes, including the trematode *S. mansoni*, and cyclophyllidean cestodes *E. multilocularis* and *H. microstoma*, despite biochemical evidence to the contrary from the related *H. diminuta* and *Taenia crassiceps*^{339,340}. In the case of *S. mansoni*, although putative

β -oxidation genes were previously identified in the genome¹⁸², they belonged to Compara families that lack predicted β -oxidation genes from other species. However, experiments that are consistent with (but do not directly demonstrate) β -oxidation occurring during *S. mansoni* egg production have been reported⁹². Thus, it is likely that *S. mansoni* and the cyclophyllidean tapeworms (*Hymenolepis*, *Taenia*, *Echinococcus*) have highly diverged β -oxidation genes; indeed, these species (except for *T. solium*) were excluded from the Compara families containing β -oxidation genes, but did share some Pfam domains with the latter. Schistosomes and cyclophyllidean tapeworms live in the glucose-rich environments of the blood and small intestine, respectively, and have transporters for glucose uptake^{341,342}, so they may have diverged to adapt to using glucose and glycogen as their main energy sources rather than lipids.

Before β -oxidation, fatty acids are linked to coenzyme A to form acyl-CoAs (activated fatty acids) by acyl CoA synthetase (EC 6.2.1.3). We found a gene family of acyl CoA synthetases that had expanded in clade IVa nematodes (family 150310; 7-10 copies per *Strongyloides* species versus 2 in *C. elegans*; **Supplementary Fig. 22i**). This was consistent with β -oxidation in infective larvae and free-living adults of *Strongyloides ratti*³³⁶. Also consistent with this was an expansion in clade IVa of a family containing *C. elegans* *trcs-1* and human AADAC, a putative arylacetamide deacetylase and microsomal lipase, predicted to break down triglyceride to fatty acids (family 167895; 9-10 per *Strongyloides* species; **Supplementary Fig. 22j**).

Ketone bodies

Acetyl CoA produced by β -oxidation, glycolysis or glycogen breakdown is usually fed into the citric acid cycle. However, under starvation conditions, some tissues (the liver in humans) convert oxaloacetate into glucose (gluconeogenesis), rather than using it as an acceptor to bring acetyl CoA in to the citric acid cycle. Under these conditions, acetyl CoA is converted into acetoacetate, D-3-hydroxybutyrate and acetone, three water-soluble compounds known as 'ketone bodies'. These water-soluble compounds are transported to key tissues (the brain in humans) for conversion back to acetyl CoA, to fuel the citric acid cycle. We found that the KEGG pathway for synthesis and degradation of ketone bodies had significantly lower coverage in filaria compared with other nematodes (**Supplementary Fig. 21a, Supplementary Table 18f**). In particular, the hydroxymethylglutaryl CoA cleavage enzyme (EC 4.1.3.4), involved in conversion of acetyl CoA to acetoacetate, and 3-ketoacid CoA transferase (EC 2.8.3.5) that converts ketone bodies back to acetyl CoA, appeared to be missing from filarial species (but not their outgroup in clade III, *Dracunculus*; **Supplementary Table 19b**). Thus, filarial worms seemed to have lost the ability to make and use ketone bodies (**Fig. 5b**), perhaps because, unlike most other helminths studied here (except *Trichinella* and *Bursaphelenchus*; **Supplementary Table 12**), they are tissue parasites that do not have a free-living stage and are less likely to suffer periods of starvation.

4.3 The glyoxylate cycle: linking lipid and carbohydrate metabolism

Absent from most metazoans other than nematodes⁸⁹, the glyoxylate cycle (M00012; **Fig. 5a**) bypasses multiple decarboxylation steps of the citric acid cycle to directly cleave isocitrate into

glyoxylate and succinate using the enzyme isocitrate lyase (ICL; EC 4.1.3.1). In a subsequent step, malate synthase (MS; EC 2.3.3.9) is used to synthesise malate from glyoxylate and acetyl CoA (**Fig. 5b**). The ICL/MS enzymes are encoded by a fusion gene in nematodes that phylogenetic analyses suggest was gained by the common ancestor of nematodes, from bacteria⁸⁹. In nematodes, the glyoxylate cycle provides a way to convert lipids via acetyl CoA into glucose (via gluconeogenesis) and other carbohydrates (e.g. trehalose), and is active during embryogenesis of *Ascaris lumbricoides* eggs, at a time when the rate of β -oxidation and carbohydrate (glycogen and trehalose) synthesis are also greatest^{263,343,344}. Glycogen and trehalose can potentially serve as energy reserves required for larvae to migrate in the host body, or during anhydrobiosis³⁴⁵. In some free-living and plant parasitic species, it has also been proposed that the glyoxylate cycle may not function to convert lipid to carbohydrate, as in *Ascaris* eggs, but instead converts ethanol into glucose or glycogen³⁴⁶. We found the corresponding KEGG module (M00012) in most nematodes (with notable absence in many clade III nematodes, including many filaria), but not platyhelminths (**Supplementary Table 18c; Fig. 5a**), although biochemical evidence supports its presence in *Fasciola*³⁴⁷. ICL and MS, the two enzymes unique to this cycle, were absent from filarial species (although they have previously been detected biochemically in *Setaria digitata*³⁴⁸, which appears to have some unusual biochemistry but was not among the worms studied in this work) and absent from *Trichinella* species (**Supplementary Table 19c**). However, the glyoxylate cycle genes were found in *Dracunculus*, the outgroup of filaria within clade III. Both filaria and *Trichinella* are tissue parasites that lack free-living stages, so may not need the metabolic diversity required to survive starvation conditions, such as the ability to convert lipid to carbohydrate.

We found an additional nematode enzyme that flatworms lack: alanine-glyoxylate transaminase (EC 2.6.1.44) that converts glyoxylate and alanine to pyruvate and glycine (**Supplementary Table 18a, Fig. 5b**). In plants, some of the glyoxylate produced (from glycolate) during photorespiration is converted to glycine by alanine-glyoxylate transaminase, and some is oxidised by glycolate oxidase (EC 1.1.3.15) to oxalate³⁴⁹. Similarly, some bacteria convert glyoxylate from the glyoxylate cycle into glycine using alanine-glyoxylate transaminase³⁵⁰. Glycine (plus tetrahydrofolate, THF) can then be converted to 5,10-methylenetetrahydrofolate by the glycine cleavage system, to provide a one-carbon pool for cellular biosynthesis (**Fig. 5b**). We found alanine-glyoxylate transaminase was absent from filaria (although present in their outgroup in clade III, *Dracunculus*) and *Trichinella* (**Supplementary Table 18a**), consistent with our finding that filaria and *Trichinella* had lost the glyoxylate cycle, and that filaria had also lost key genes required for the glycine cleavage system (see *Amino acid metabolism* below). Thus, we propose that most nematodes, except filaria and *Trichinella*, convert some of the glyoxylate from the glyoxylate cycle to glycine, as input into the glycine cleavage system.

4.4 Carbohydrate metabolism

Scavenging carbohydrates from host food

Intestinal helminths such as tapeworms, some ascarids, and some hookworms such as *Oesophagostomum*, are thought to obtain most nutrients from intestinal contents of their hosts (**Supplementary Table 12**). We found that a family of α -glucosidases (EC 3.2.1.20; CAZyme class GH31), enzymes that break down starch and disaccharides (e.g. sucrose) to glucose, had independently expanded in clade IIIb (Ascaridomorpha) and clade Vc (family 69769; 51% predicted secreted; 19-31 copies per *Ascaris* species; 7-22 in *Ancylostoma*; **Supplementary Fig. 22a**). Excretory-secretory products of *A. suum* include many GH31 family glycosyl hydrolases, which include α -glucosidases³⁵¹. Thus, α -glucosidase genes may be secreted to enable the worms to break down host food and so use it as an energy source. Among hookworms, the most family members were found in *Oesophagostomum* (31), but it was also expanded in species thought to feed principally on blood (**Supplementary Table 12**) such as *Ancylostoma* species (7-22 copies), suggesting that they also obtain some energy from host food.

Malate dismutation

Many adult helminths live in low-oxygen conditions within host tissues or the host intestine and rely at least partially on anaerobic metabolism to produce ATP, including the malate dismutation pathway^{93,352-354}. In this pathway, glycolysis converts carbohydrates to phosphoenolpyruvate (PEP) that is then carboxylated by phosphoenolpyruvate carboxykinase (PEPCK) to form oxaloacetate and subsequently reduced to malate. The malate is partly oxidised to acetate and partly reduced to succinate/propionate (**Fig. 5b**), which are excreted as end products⁹³. We found that a PEPCK family had expanded in clade IIIb (Ascaridomorpha; family 146234; EC 4.1.1.32; 7-9 copies in *Ascaris* spp. versus 2 *B. malayi*, 2 *O. volvulus*; **Supplementary Fig. 22c**). The extra copies of PEPCK may possibly enable ascarids to channel phosphoenolpyruvate into the malate dismutation pathway, rather than into aerobic metabolism (citric acid cycle), as malate dismutation is thought to be a major source of energy in *Ascaris*^{355,356}. Alternatively, PEPCK is required by the glyoxylate cycle, which converts acetyl CoA to carbohydrates during embryogenesis of *Ascaris* eggs²⁶³. We also saw a small expansion of another enzyme involved in malate dismutation, methylmalonyl CoA epimerase (EC 5.1.99.1) in clade IIIb (family 586945; 2-4 genes per species in this clade but single-copy elsewhere; **Supplementary Fig. 22e**). In addition, an intracellular cobalamin (vitamin B12) trafficking chaperone gene family, containing *C. elegans cblc-1* and human *MMACHC*³⁵⁷⁻³⁵⁹, had expanded in clade IIIb (family 646695; 4 *A. suum* versus 1 *C. elegans*; **Supplementary Fig. 22d**). This may allow increased mitochondrial uptake of vitamin B12, which is required for malate dismutation in mitochondria³⁵³, and absorbed in large amounts by some helminths^{93,360}. Surprisingly, we also found a CobQ/CbiP cobyrinic acid synthase family in clade IIIb (e.g. *A. suum* ASU_02377; also in the outgroups *Amphimedon* (sponge) and *Capitella* (annelid); family 1160166; EC 6.3.5.10), an amidating enzyme involved in cobalamin synthesis that converts adenosylcobyrinic acid to adenosylcobyrinic acid³⁶¹. This family may have been gained by horizontal

transfer from bacteria, since it had CbiA (Pfam PF01656) and GATase_3 (PF07685) domains, a combination almost exclusively found in bacteria and archaea.

Phylogenetic analysis confirmed that these ascarid proteins form a cluster within a clade of Bacteroidales sequences, and probably represent a genuine lateral gene transfer from this group (**Supplementary Fig. 23a**). The phylogeny confirmed that the other invertebrate sequences (from *Amphimedon* and *Capitella*) were related to a different bacterial group (γ -proteobacteria) and seemed likely to be contaminants rather than genuine lateral gene transfers. That is, the *Amphimedon* gene was present in the v1.0 assembly for that species used in our Compara database, but seemed to be deleted in v2.0. Both this gene and the *Capitella* sequence were on small contigs (~10 kb and ~7 kb, respectively) with only single-exon genes or genes annotated with only short introns, and with strong BLAST hits to the genomes of a group of marine bacteria that are frequently found associated with marine invertebrates.

Bacteroides are common in the mammalian intestine where *Ascaris* is found³⁶². This nematode enzyme probably has a different role in clade IIIb than in bacteria, perhaps in converting scavenged vitamin B12 and B12-like compounds into forms such as adenosylcobalamin that are useful for malate dismutation. Indeed, in some archaea, it appears that adenosylcobyrinic acid is the entry point for adenosylcobalamin salvage³⁶³, although this requires additional enzymes such as CbiB that we did not find in *Ascaris*. The *A. suum* member of the family (ASU_02377) is overexpressed in the *Ascaris* intestine³⁶⁴, which would be consistent with a role in salvage from host tissue or host intestinal contents.

Duplications (defined here as EC numbers that are represented by ≥ 2 genes in ≥ 2 of the 81 nematode and platyhelminth species) were identified in the 'glyoxylate and dicarboxylate' KEGG pathway (which overlaps with malate dismutation; **Supplementary Table 20j**), with examples of paralogues for 92% its enzymes. For example, there were duplications in the malate dismutation enzyme propionyl-CoA carboxylase (EC 6.4.1.3), which converts D-methyl-malonyl CoA to propionyl CoA, in many nematodes and in trematodes (**Supplementary Table 20j**).

Propionate shunt

Acetate (via acetyl CoA) and propionate are common end products of malate dismutation but some helminths may re-metabolise a portion of them under certain conditions. In clade Va (strongylid-other), we saw an expansion of a family containing *C. elegans* dehydrogenase *alh-8*, involved in a cobalamin-independent 'propionate shunt' pathway for converting propionate to acetyl CoA⁹⁴ (family 476273; EC 1.2.1.18/1.2.1.27; 5-8 copies per *Haemonchus* species; 13 *Teladorsagia circumcincta*; **Supplementary Fig. 22f, Fig. 5b**). One possibility is that *Haemonchus* and close relatives use this pathway to convert propionate from malate dismutation, or possibly residual unabsorbed propionate from bacterial fermentation in the ruminant host⁹⁵, to acetyl CoA that can be converted to acetate (and thereby produce additional ATP)³⁶⁵. The *C. elegans* propionate shunt genes are predicted to be localised to the mitochondria⁹⁴, where malate dismutation occurs. The propionate shunt may be more important in nematodes than flatworms, since the 'beta-alanine' KEGG pathway (which overlaps with the propionate shunt) included two enzymes that were found

in nematodes but not flatworms: 4-aminobutyrate aminotransferase (EC 2.6.1.19, which interconverts β -alanine and malonic semialdehyde, a propionate shunt intermediate) and the propionate shunt enzyme acyl-CoA dehydrogenase (EC 1.3.8.7, corresponding to *C. elegans* *acdh-1*⁹⁴; **Supplementary Table 18a**; **Supplementary Table 20f**).

The lactate dehydrogenase pathway

The lactate dehydrogenase (LDH) pathway is an alternative pathway for the anaerobic production of ATP from carbohydrates³⁵⁴. The LDH pathway produces only two molecules of ATP per molecule of glucose, compared with three for the malate dismutation pathway, which may be why *Ascaris* prefers the malate dismutation pathway to survive in the fluctuating carbohydrate supply of the host intestine²⁶³. In contrast, some adult flatworms such as *Schistosoma* spp. (which live in glucose-rich blood), *C. sinensis*, *E. granulosus*, and *Taenia* spp. are known to use the LDH pathway³⁵⁴. In agreement with this, the Compara family that contains *LDH* (EC 1.1.1.27) has a higher copy number in cestodes and schistosomatids (e.g. 6 *E. multilocularis*, 3 *S. mansoni*), compared with most nematodes (e.g. 1 *A. suum*; family 136770; **Supplementary Fig. 22g**). An exception was the *Ancylostoma* clade, which had many *LDH* genes (7 *A. duodenale*, 8 *A. caninum*, 6 *A. ceylanicum*). The LDH pathway may therefore be a predominate source of ATP in *Ancylostoma* parasites that, like schistosomes, feed on glucose-rich blood for part of their life cycle.

The GABA shunt

In clade IIIb (Ascaridomorpha), we saw a small expansion of a glutamate dehydrogenase family (family 114732; EC 1.4.1.2, EC 1.4.1.3 or EC 1.4.1.4; 4-5 per *Ascaris* species; **Supplementary Fig. 22h**). Glutamate dehydrogenase interconverts glutamate and α -ketoglutarate, and in this case may possibly be involved in conversion of α -ketoglutarate to succinate via glutamate and gamma-aminobutyric acid (GABA), known as a GABA shunt (**Fig. 5b**). Although the other enzymes involved in the GABA shunt (GABA- α -ketoglutarate aminotransferase, EC 2.6.1.19; glutamate decarboxylase, EC 4.1.1.15; and succinate-semialdehyde dehydrogenase, EC 1.2.1.24) appeared to be present in some nematodes, they were missing from flatworms (in *Clonorchis* a gene was annotated as glutamate decarboxylase but had a top BLAST hit to bacteria, indicating possible sequence contamination; **Supplementary Table 19d**). Based on the existence of glutamate dehydrogenase, glutamate decarboxylase and aminotransferase, in *Ascaris*, and its excretion of GABA under anaerobiosis, it was previously suggested that it may have a GABA shunt (see review³⁶⁶). In humans, the GABA shunt (also known as '4-aminobutyrate bypass') may regulate redox homeostasis by controlling intracellular levels of α -ketoglutarate and fumarate³⁶⁷. Indeed, glutamate dehydrogenase has been reported in the tapeworm *Hymenolepis diminuta* and in blood feeding stages of *Haemonchus contortus*, in which it is proposed to maintain the NAD/NADH redox balance in response to changes resulting from anaerobic metabolism (malate dismutation and LDH pathways)^{250,368,369}. It is curious that we found glutamate dehydrogenase but not the other GABA shunt enzymes in flatworms such as *Hymenolepis* (**Supplementary Table 19d**). Furthermore, while glutamate dehydrogenase is present in the mitochondria of mammals, the *Hymenolepis* and

Haemonchus proteins appear to be cytoplasmic^{368,369}. An alternative role for glutamate dehydrogenase is to convert glutamate to α -ketoglutarate (i.e. the reverse of the reaction that occurs in the GABA shunt) when amino acids are needed as precursors for glucose synthesis (gluconeogenesis), or for energy (the α -ketoglutarate can be used in the citric acid cycle)³⁴⁹, and this may possibly be its main role in flatworms.

Glycogen synthesis

In trematodes, we identified an expansion of a CAZyme class GT8 glycosyltransferase family, especially in *Clonorchis* (family 476359; 14 *Clonorchis*, 4 *S. mansoni*; **Supplementary Fig. 22b**). Expansion of GT8 enzymes (involved in lipopolysaccharide and glycogen biosynthesis) in trematodes may be linked to the reliance of their cercariae (free-living larvae) on energy stored within glycogen granules^{10,304}.

4.5 Amino acid metabolism

Glycine cleavage system

The GCS uses four proteins (T-, P-, L- and H-proteins) to catabolise glycine when present in high concentrations, to form a one-carbon donor (5,10-methylenetetrahydrofolate) for various biosynthesis reactions (**Fig. 5b**), including purine, thymidylate and methionine synthesis³⁷⁰. In at least some bacteria, the GCS can be run in reverse to synthesise glycine. Though present in many helminths, the absence of detectable T- and P- proteins (EC 2.1.2.10 and 1.4.4.2, respectively) suggests that the GCS is missing from cestodes and filarial worms (**Supplementary Table 19e**). In contrast, the T- and P-proteins appeared to be present in *Dracunculus*, sister group to filaria in clade IIIc, as it had members in the same Compara families as *C. elegans* T-protein (*gcst-1*) and P-protein (*glde-1*). EC 1.8.1.4 (L-protein) is important for many other processes and seemed to be ubiquitous. H-protein has been predicted to be present in the cestodes *E. multilocularis* and *E. granulosus*^{13,371} and the filarial nematode *B. malayi*²². Indeed we found members from the latter species in the same Compara family as *C. elegans* H-protein (*gcsh-1*). In cestodes and filarial nematodes, the H-protein could therefore represent either a vestigial gene from the lost pathway or perform an alternative and uncharacterised function. Indeed, 'partial glycine cleavage systems' have been found in the protozoan parasites *Trichomonas vaginalis*, which is anaerobic (just has L-protein and H-protein³⁷²), and *Spironucleus salmonicida* (has H-protein³⁷³), which is adapted to micro-aerobic environments. If cestodes and filaria cannot make 5,10-methylenetetrahydrofolate (to create a one-carbon pool) from glycine (plus tetrahydrofolate, THF) using the glycine cleavage system, they may instead catabolise serine (plus THF) to form glycine and 5,10-methylenetetrahydrofolate using the enzyme serine hydroxymethyltransferase (EC 2.1.2.1; **Supplementary Table 18a**).

Duplications were identified in the KEGG glycine-serine-threonine (GST) KEGG pathway, with examples of paralogues for 71% of its enzymes (**Supplementary Table 20b**). For example, there were lineage-specific duplications in the enzymes that convert serine into glycine (serine

hydroxymethyltransferase, EC 2.1.2.1; duplicated in trematodes and clade V), and produce serine from pyruvate (EC 4.3.1.19; duplicated in clades IV and V).

Amino acid auxotrophies

Parasitic nematodes and platyhelminths have a wide host range and their nutritional requirements are expected to vary. Numerous biochemical and mineral components are essential nutrients, hence we examined the potential impact of the differences in annotated enzymes on the biosynthetic capacity (using Pathway Tools, see **Methods 13**) by predicting amino acid and vitamin auxotrophies for all tier 1 species. Nematodes and platyhelminths were predicted to be auxotrophic for eight of the nine amino acids essential to humans and most animals (Phe, His, Lys, Leu, Met, Thr, Trp, and Val) (**Supplementary Table 18g**). Among these, valine was predicted by Pathway Tools to be auxotrophic in all nematodes and platyhelminths except *C. sinensis*; however closer examination, and in the absence of experimental data, indicated that this might be due to gene model errors in this published genome¹¹. Auxotrophy of isoleucine, another amino acid essential for most animals³⁷⁴, was not confidently predicted by Pathway Tools for many nematodes. However, isoleucine auxotrophy is highly likely, given that nematodes and platyhelminths were predicted to be auxotrophic for leucine and that the pathways for producing isoleucine and leucine are similar, except for the threonine to 2-oxobutanoate step (EC 4.3.1.19). Besides these nine amino acids, arginine was also predicted to be an essential amino acid for all species investigated here (**Supplementary Table 18g**). This has been recognised previously in nematodes³⁷⁵ but not in platyhelminths, and was due to loss of the urea cycle enzymes necessary for converting the precursor ornithine to arginine (ornithine carbamoyltransferase, argininosuccinate synthase, argininosuccinate lyase; **Supplementary Table 18a**). Unlike platyhelminths and nematodes, the arginine synthesis pathway has been conserved in humans and other vertebrates.

4.6 Nucleotide metabolism

Pyrimidine synthesis

The KEGG pathway for 'alanine, aspartate and glutamate metabolism' lacked both carbamoyl phosphate synthetase (EC 6.3.5.5; CPS-II) and aspartate transcarbamylase (EC 2.1.3.2; ATCase), the first two enzymes in pyrimidine synthesis, across all cestodes analysed here (**Supplementary Table 19f**). We do however note that ATCase has been reported in the cestodes, *Moniezia benedeni* and *Hymenolepis diminuta*³⁷⁶, and CPS-II has in *H. diminuta*, albeit with low activity³⁷⁷. Furthermore, indirect evidence from *in vitro* studies suggests that *M. corti* larvae can synthesise pyrimidines *de novo*³⁷⁸. In addition, some cestode genes had the Pfam domains associated with CPS-II (PF02786, PF02787), but not the domain usually found in ATCase (PF02729). These two enzymes were present as a fusion gene in *C. elegans* (*pyr-1*) and *S. mansoni* (*Smp_186670*). Their absence suggests that many cestodes might be lacking pyrimidine *de novo* biosynthesis. In agreement with this, previous analysis of the *Echinococcus granulosus* genome led to the

suggestion that it cannot synthesise pyrimidines³⁷¹. Thus, some cestodes may instead depend on salvage from their hosts, which may occur even in helminth species that can synthesise pyrimidines³⁴⁶. Loss of one or more enzymes in the pyrimidine synthesis pathway has also been reported in several filaria and *T. spiralis*²⁴. Both cestodes and filaria were missing KEGG module M00051 that is part of pyrimidine metabolism (**Fig. 5a**). Indeed we did not find the enzymes ATCase or CPS-II in any of the filaria (except for predicted CPS-II genes in *B. timori* and *D. immitis*, whose top BLASTP hits to UniProt indicated they were bacterial contaminant or endosymbiont genes, respectively). It has previously been shown that the *Wolbachia* found in *D. immitis* has pyrimidine synthesis genes, and suggested that many filaria depend on their endosymbionts for pyrimidine synthesis^{23,379}. In contrast, we did find evidence of pyrimidine metabolism in *Dracunculus*, the outgroup in clade III that does not contain an endosymbiont (**Supplementary Table 19f**).

Purine synthesis

Purine synthesis starts with the activation of ribose phosphate to phosphoribosyl pyrophosphate (PRPP), a step that was present in all species examined (KEGG Module M00005, **Supplementary Table 18d**). Despite some indirect biochemical evidence to the contrary in *Brugia pahangi* and *Dirofilaria immitis*³⁸⁰, the synthesis of IMP from PRPP (represented by module M00048) was completely absent from parasitic platyhelminths, filaria and most clade IV nematodes (**Fig. 5a**). This agreed with previous analyses of genomes of filaria such as *B. malayi*, *D. immitis* and *L. loa*²²⁻²⁴, and of flatworms such as *Echinococcus granulosus*³⁷¹ and *Schistosoma japonicum*⁹. Many platyhelminths and nematodes clearly need to salvage purines, consistent with previous reports⁹¹. It has been suggested that in some filarial species such as *B. malayi* and *D. immitis*, *de novo* purine biosynthesis is provided by their *Wolbachia* endosymbionts^{22,23}. Although complete pathways could be detected for several species (including *Trichuris* spp., *C. elegans*, *Dracunculus medinensis* and *Haemonchus contortus*), for most of the other nematodes of clades I, IIIb and V, a near-complete pathway could be detected, lacking the terminal two steps (AICAR transformylase, EC 2.1.2.3, and IMP cyclohydrolase, EC 3.5.4.10; **Supplementary Table 19g**) that are needed to synthesise IMP from a precursor ribonucleotide. This suggests that these latter enzymes are particularly difficult to identify in nematodes, have been lost from several species, or that an alternative mechanism to produce IMP is used. In two cases (*Elaeophora* and *Rhabditophanes*), these terminal steps were the only parts of the pathway to be detected, although the *Elaeophora* gene encoding these steps (*EEL_0000649601*) appeared to be a contaminant as it had its top UniProt BLASTP hit to bacteria (to UniProt J2VZF7 from *Phyllobacterium*, *E*-value 0.0, 89% id.). *Fasciola hepatica* appeared as an outlier in our analyses, with a complete pathway apparently present. However, inspection of individual hits revealed that the pathway was provided by the recently reported *Neorickettsia* endosymbiont¹².

Enzymes required for interconversions between IMP and other purine nucleosides (represented by KEGG modules M00049 and M00050) were ubiquitous in nematodes and platyhelminths. They were detected in all tier 1 species except for *F. hepatica*, which lacked EC 4.3.2.2 (adenylosuccinase) and *H. contortus*, which appeared to lack EC 6.3.5.2 (GMP synthase), until a

reasonable *Fasciola* candidate (D915_00696) was obtained by looking at Compara families and a *Haemonchus* candidate was found from a BLASTP search of the published Australian strain²⁵⁰ using *H. placei* GMP synthase as a query sequence (**Supplementary Table 18a-b**).

4.7 Metabolism of cofactors and vitamins

Haem metabolism

The only KEGG pathway where platyhelminths had significantly higher coverage than nematodes was 'porphyrin and chlorophyll metabolism' (**Supplementary Fig. 21a**). This was primarily due to presence of multiple enzymes of the haem biosynthesis pathway in platyhelminths, while all but the terminal ferrochelatase (EC 4.99.1.1) step were missing from most nematodes (**Supplementary Table 19h; Supplementary Table 20i**), consistent with their inability to synthesise haem^{47,381,382}. Platyhelminths, including *S. mansoni* (despite conflicting biochemical data⁴⁷), appeared to have a complete haem biosynthesis pathway, although some parts of the pathway are quite diverged. For example, our EC annotation (**Supplementary Table 18a**) was missing URO synthase (EC 4.2.1.75) and protoporphyrinogenase (ECs 1.3.3.4/1.3.5.3), but we were able to find reasonable candidates for these genes in platyhelminths by looking at Compara families and BLAST searches. That is, the Compara family containing human URO synthase (*UROS* or *HemD*, *ENSG00000188690*) was family 902846, which included the *S. mansoni* gene *Smp_079840*, a good candidate for *S. mansoni* URO synthase. In addition, using human protoporphyrinogenase (*PPOX*) in a BLASTP search against WormBase ParaSite¹⁶⁰, we found a hit to *S. mansoni* *Smp_068250* (*E*-value 1e-6, 38% identity), which was the only *S. mansoni* gene that had a protoporphyrinogenase domain (IPR004572; *E*-value 3e-44), so was a good candidate for *S. mansoni* protoporphyrinogenase. In support of the presence of haem biosynthesis in platyhelminths, transcripts for haem biosynthesis enzymes were previously reported in schistosome EST data sets³⁸³, while haem biosynthesis has been reported in the free-living platyhelminth *Schmidtea mediterranea*³⁸⁴.

Vitamin metabolism

Retinol (vitamin A) can be obtained from the diet either by hydrolysing dietary retinyl esters (e.g. by using EC 3.1.1.3, present in most nematodes and many flatworms; **Supplementary Table 18a**) or splitting carotenoids (e.g. β -carotene). The latter route uses beta-carotene dioxygenase (EC 1.13.11.63, formerly EC 1.14.99.36), which was only present in some nematodes and was absent from flatworms (**Supplementary Table 18a-b**); and retinol dehydrogenase (EC 1.1.1.300), which appeared to be present in most nematodes and platyhelminths, since they had members belonging to a family (id. 133629) containing the human retinol dehydrogenase gene (*ENSG00000139988* = *RDH12*; **Supplementary Table 18b**). Further, like humans, most nematodes and platyhelminths were predicted to salvage thiamine (vitamin B1)³⁸⁵ and pyridoxal (vitamin B6)³⁸⁶, two important cofactors (**Supplementary Table 18g**).

Folate derivatives (e.g. tetrahydrofolate) are important co-factors facilitating one-carbon transfer in key biosynthetic processes (synthesis of purines, pyrimidines, methionine) and appear to be procured from the environment³⁸⁷ as folic acid, as is the case for humans and other animals³⁸⁸. Folate polyglutamylation, a key step for generating the preferred form of the cofactor utilised by enzymes (by enzymes dihydrofolate synthetase, EC 6.3.2.12, and folypolyglutamate synthetase, EC 6.3.2.17), was predicted to be active in all nematodes (**Supplementary Table 18g**); its apparent absence in platyhelminths was likely due to the challenge of annotating divergent enzymes³⁸⁷, since we did find putative flatworm folypolyglutamate synthetase genes (EC 6.3.2.17) by analysing Compara families (**Supplementary Table 18b**). In general, spiruromorphs (including filaria) and cestodes seemed to lack many pathways for post-processing of folates (**Supplementary Table 18g**).

5. Identifying potential new anthelmintic drug targets and drugs

Authors: Avril Coghlan, James Cotton, Andrew R. Leach, Prudence Mutowo, Jane Lomax and Noel O'Boyle

5.1 Summary of approach

There is a pressing need for new anthelmintic drugs, since existing ones suffer from low efficacy, serious side effects or rising drug resistance in parasite populations³⁸⁹⁻³⁹¹. The availability of genome data provides new opportunities to identify compounds structurally different from existing anthelmintics, that may provide new treatments. This can be done by repurposing existing drugs and/or by identifying biologically active compounds from other areas of drug discovery as starting points for hit/lead discovery. We therefore searched for the closest sequence (BLASTP) match of each helminth protein in the ChEMBL database¹⁷⁷, and if a significant match was found, we retrieved compounds with bioactivities for that ChEMBL target (**Methods 14**). The helminth protein was then assigned a score based on its attractiveness as a potential new drug target, taking into account factors such as the strength of the BLAST match, whether the ChEMBL target lacked human homologues, and whether *C. elegans* or *D. melanogaster* homologues had lethal or sterile phenotypes when disrupted. This scoring system was devised using a training set of known/likely targets of existing anthelmintic drugs used for humans, to ensure that helminth proteins having high scores were good candidates for new anthelmintic drug targets. Our goal was to identify candidates for new classes of anthelmintic compounds, rather than molecules structurally similar to known anthelmintic compounds. Molecules associated with known targets are most attractive for drug development if, for example, the target has a lethal knockout phenotype. Thus, we focussed on compounds active against our top (15% of) highest-scoring targets. We then filtered the target-compound pairs to retain those that were the most promising by virtue of having a high pChEMBL score (reflecting high potency/affinity for the ChEMBL target) or appearing in a PDB structure with

the ChEMBL target. This gave us a set of top candidate molecules, which included some approved drugs (mostly not anthelmintics). These could be considered for repurposing as novel anthelmintics, which would save considerable effort and expense, a key factor in drug discovery for neglected diseases. The other compounds in the set are typically from the medicinal chemistry literature and so represent molecules to test for anthelmintic activity.

ChEMBL is a manually curated database of molecules with biological or pharmacological activity. To establish the suitability of ChEMBL as a source of potential new anthelmintic agents, we checked for the presence of existing compounds. Of 261 known anthelmintic compounds (current veterinary or medical anthelmintics, plant-nematicidal agents, or medicinal chemistry compounds) 208 out could be identified, including 23 of the 24 drugs used for humans with WHO ATC code P02 (WHO anthelmintics) and 33/37 veterinary drugs with the WHO ATCvet classification (**Supplementary Table 21a**, columns A-B). A list of known/likely targets for the 24 WHO compounds was collated by searching the DrugBank database³⁹² and by additional literature searches. At least one plausible single-protein target was identified in ChEMBL for 19/24 WHO anthelmintics, and for an additional nine known anthelmintic compounds not in the WHO set (**Supplementary Table 21b**). Other WHO compounds lacked targets in ChEMBL either because their target proteins were unknown (e.g. niclosamide, desaspidin, dichlorophen); or their known/likely targets lacked sequence matches in ChEMBL (e.g. oxamniquine); or their targets were complexes and not single proteins in ChEMBL (e.g. oxantel). The suitability of ChEMBL for searching was underlined by the fact that ChEMBL contained homologues for the majority of known/likely single-protein targets of existing anthelmintic drugs.

5.2 Ranking helminth proteins as likely drug targets

To identify helminth proteins likely to be good potential drug targets, we ran BLASTP to compare all (528,469) helminth proteins from the 33 tier 1 species (those with high-quality assemblies; **Methods 2**) with known targets from the ChEMBL 21 database that contains 1,592,191 compounds and 11,019 targets. We focussed on the 6,261 single protein ChEMBL targets that may be easier to develop as drug targets, rather than protein complexes. 106,278 helminth genes were found with top BLASTP hits ($E \leq 1e-10$) to 3,994 individual single-protein ChEMBL targets. We then assigned a target score to each of the 106,278 helminth proteins in an attempt to quantify their likely quality as targets for known compounds (**Methods 14**). The major contributors to this score were the quality (E -value and target coverage) of the BLASTP match between the helminth and ChEMBL proteins; whether the ChEMBL target had a close human BLAST match (since targeting a protein that lacks a human homologue is less likely to cause undesirable side-effects); and whether the helminth gene had a *C. elegans* or *D. melanogaster* homologue with a relevant phenotype (for example, a lethal or sterile phenotype) (**Methods 14**). Other minor contributors included whether the helminth protein was a predicted chokepoint enzyme; whether it was expressed in key life cycle stages (for example, adult); whether it had homologues in most members of a major helminth clade (for example, in most nematodes); whether it lacked within-species paralogues; whether it belonged to a Compara family with a highly conserved alignment;

whether the matching ChEMBL protein had a structure in the PDB¹⁷⁹; and whether it was from a non-chordate animal. The list of 106,278 helminth genes contained many related genes (for example, 220 tubulin genes). To remove this redundancy, we first took only the highest-scoring helminth match to each ChEMBL protein, and then further filtered to take only the highest-scoring helminth member from each Compara family, leaving 1,925 helminth proteins (matching 1,925 ChEMBL targets).

5.3 Choice of weightings in the scoring system for targets

When developing our scoring system for helminth proteins, we initially assigned weights to contributors based on prior knowledge about favourable properties (e.g. a strong BLASTP match to a ChEMBL target and a predicted lethal knockout phenotype were assigned large weightings). We then checked that the known/likely targets for most WHO anthelmintics (listed in **Supplementary Table 21b**) were in the top 15% (top 289/1925) of this ranked list, and manually adjusted the weightings until this was the case. In this way, the known/likely targets of existing anthelmintic drugs used for humans were used as a ‘training set’ to ensure that high scores reflected an increased likelihood of being good candidate targets for new anthelmintics. Using our final set of weightings (**Methods 14**), the top 15% of our target list included at least one known/likely target for 17/19 WHO anthelmintics with targets in ChEMBL (ivermectin, praziquantel, albendazole, mebendazole, metrifonate, bithionol, stibophen, pyrantel, levamisole, bephenium, niridazole, triclabendazole, tiabendazole, ciclobendazole, flubendazole, fenbendazole, pyrvinium). The two exceptions were piperazine and diethylcarbamazine. Single-protein targets for these compounds were present in ChEMBL and had (low-ranked) hits in our analysis, but these may not be the key targets *in vivo*^{393,394} (**Supplementary Table 21b**). Finally, we checked whether 13 potential drug targets, identified in at least two previous helminth genome studies^{17,176,275,395}, tended to score well in our scoring system (**Supplementary Table 21c**). Eleven were in the list of helminth targets, eight of which were in the top-scoring 15% of targets, and two in the top-scoring 25%. Our scoring approach thus captured known targets of most clinical anthelmintics, and was able to identify candidates from previous *in silico* studies, giving us confidence that other helminth proteins that have high scores are good candidates for new anthelmintic drug targets.

5.4 Identifying potential new anthelmintic targets, and drugs

From the BLASTP searches against ChEMBL single-protein targets using the helminth proteomes as queries (see above), we identified 827,889 small molecules with recorded biological activities. Concentrating on the top 15% (289 of 1925) highest-scoring helminth proteins (**Supplementary Table 21d** column A), we took compounds associated with their top BLASTP hits in ChEMBL and compounds associated with the top BLAST hits of other members of the same Compara family. Of our 289 highest-scoring helminth targets, 286 targets had such compounds (292,499 unique compounds). To increase the chance that target-compound associations in ChEMBL corresponded to bioactivity *in vivo*, we used two filters to select the target-compound pairs with the most

convincing evidence: (i) co-appearing in a PDB¹⁷⁹ structure with the ChEMBL target; or (ii) having a high pChEMBL score (median >5 for phase III/IV drugs, or >7 for medicinal chemistry compounds), reflecting high potency/affinity for the ChEMBL target). Since existing phase III/IV drugs are more attractive for developing (repurposing) as novel anthelmintics, we used a less stringent pChEMBL filter for phase III/IV drugs (median pChEMBL >5) than for medicinal chemistry compounds (median >7). This gave us a list of 52,154 'top candidate molecules', which included 795 approved drugs (phase IV) and 164 phase III drugs that have 102 targets. These drugs, which could be considered for repurposing as novel anthelmintics, were classified into 730 chemical classes (**Methods 14**). The remaining 51,195 molecules, which were medicinal chemistry compounds (rather than drugs), spanned 13,191 additional chemical classes and represent a screening pool that the research community could test for anthelmintic activity.

5.5 Diverse screening set

In order to create a diverse screening set of compounds for screening in whole-worm assays, we first took the 6432 (12.3%) of the 52,154 compounds that were listed as commercially available in ZINC15¹⁸⁰ (**Methods 14**), which were clustered into 4543 chemical classes. Then, to select one compound from each cluster, we devised a ranking score. Contributors to this score included the QED score of the compound (a measure of oral drug-likeness³⁹⁶); the phase of approval of the compound as a drug (e.g. phase IV or III approval); whether the compound (if an approved drug) could be administered orally or topically; and whether it had serious side effects and/or predicted toxicology targets (**Methods 14**). The phase of drug approval (phase 0-IV) was given the highest weight, assuming that drugs with higher phase of approval will be quicker and cheaper to develop as novel anthelmintic drugs. For each target, the highest-scoring compound from each chemical class was taken, giving a 'diverse screening set' of 5046 compounds, including 817 drugs (with phase IV/III approval) in 661 chemical classes (**Supplementary Table 21d** columns V, W), and 4229 medicinal chemistry compounds in 3936 chemical classes (**Supplementary Table 21d** column X), excluding chemical classes containing known anthelmintic compounds (from **Supplementary Table 21a**). Note that there still were multiple compounds from some chemical classes in this screening set, because when collapsing by chemical class, we only collapsed by chemical classes for each target, so the final screening set may have compounds in the same chemical class for two different targets. A brief literature search revealed that some of our top 20 targets had candidate compounds in our screening set that have previously been shown to have anthelmintic activity: for example, a candidate for thymidylate synthase (Compara family 644442) was the anticancer drug fluorouracil, which has anthelmintic activity against *Echinococcus* larvae and cells³⁹⁷.

A self-organising map (**Fig. 6**) confirmed the diversity of these 5046 compounds, as they were distributed across 390 of the 400 cells while 236 known anthelmintic compounds were much more clustered in the map, in only 127 cells (**Supplementary Fig. 24**). Randomly selected sets of 236 of the 5046 compounds were more widely distributed than this, filling a median of 174 cells (range 147-200 for 500,000 randomisations), so a similar distribution of the novel compounds and known

anthelmintic compounds can be rejected at $P < 2 \times 10^{-6}$. The average distance between cells occupied by randomly chosen pairs of screening compounds was also significantly greater than that for pairs of known anthelmintic compounds, even when compounds sharing a single cell were excluded (median distances 9 vs 8, means 9.25 and 8.19; $P < 2.2 \times 10^{-16}$ by Wilcoxon signed-rank test for 10,000 random pairs from each group), suggesting that not only did known anthelmintic compounds share the same cell more often than expected (compared to the screening set), but that cells containing known anthelmintic compounds were also closer to each other in the map than expected.

5.6 A high priority target set, giving a smaller diverse screening set

Since 5046 compounds is still too many for smaller laboratories, we also created a second smaller screening set by identifying 40 'high priority targets' among our top 15% of targets (**Supplementary Table 21d**, column S). These consisted of several classes of interesting targets that we noticed among our top 15% of targets. Firstly they included targets lacking human homologues (three targets, as defined using our 'ChEMBL non-human score'; **Methods 14**) which suggested possible selective compounds; others were clade-specific chokepoints (for example, two targets predicted in all nematodes but no platyhelminths), so again may have specificity. Secondly, we also included targets that had greatly expanded in some helminths (five targets; from **Supplementary Table 9a**; see **Supplementary Note 2.4**), and so may perform functions important to the parasite. Thirdly, we included known/likely anthelmintic targets (from **Supplementary Table 21b**) or targets in the same pathway as known anthelmintic targets (30 targets, **Supplementary Fig. 25**), which suggested compounds that may also be anthelmintic but could possibly be more efficacious and/or give less severe side effects. These 40 high priority targets had 720 candidate compounds, including 181 drugs (phase IV/III) in 173 chemical classes, and 539 medicinal chemistry compounds in 530 chemical classes. A brief literature search revealed encouraging evidence that some of these have anthelmintic activity. For example, we suggested several compounds for the target glycogen phosphorylase, which is in the same pathway as a likely anthelmintic target (glycogen phosphorylase phosphatase, likely inhibited by niridazole), including the phase III drug alvocidib (flavopiridol) that has anthelmintic activity in *C. elegans*⁹⁶. Another example is the target cathepsin B, which had greatly expanded in clade Va (strongylid-other; **Supplementary Table 9a**), for which we suggested several compounds including the phase III drug odanacatib, which has recently been shown to have anthelmintic activity against hookworms⁹⁷.

Additional references

182. Berriman, M. *et al.* The genome of the blood fluke *Schistosoma mansoni*. *Nature* **460**, 352-8 (2009).
183. Tyson, J.R. *et al.* MinION-based long-read sequencing and assembly extends the *Caenorhabditis elegans* reference genome. *Genome Res* **28**, 266-274 (2018).
184. Delcher, A.L., Phillippy, A., Carlton, J. & Salzberg, S.L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478-83 (2002).
185. Burke, M. *et al.* The plant parasite *Pratylenchus coffeae* carries a minimal nematode genome. *Nematology* **17**, 621-637 (2015).
186. Gregory, T.R., Hebert, P.D. & Kolasa, J. Evolutionary implications of the relationship between genome size and body size in flatworms and copepods. *Heredity (Edinb)* **84** (Pt 2), 201-8 (2000).
187. Thomas, J.H. & Robertson, H.M. The *Caenorhabditis* chemoreceptor gene families. *BMC Biol* **6**, 42 (2008).
188. Coghlan, A. *et al.* nGASP--the nematode genome annotation assessment project. *BMC Bioinformatics* **9**, 549 (2008).
189. Logan-Klumpler, F.J. *et al.* GeneDB--an annotation database for pathogens. *Nucleic Acids Res* **40**, D98-108 (2012).
190. Lunt, D.H., Kumar, S., Koutsovoulos, G. & Blaxter, M.L. The complex hybrid origins of the root knot nematodes revealed through comparative genomics. *PeerJ* **2**, e356 (2014).
191. Abad, P. *et al.* Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nat Biotechnol* **26**, 909-15 (2008).
192. Castagnone-Sereno, P., Danchin, E.G., Perfus-Barbeoch, L. & Abad, P. Diversity and evolution of root-knot nematodes, genus *Meloidogyne*: new insights from the genomic era. *Annu Rev Phytopathol* **51**, 203-20 (2013).
193. Detwiler, J.T. & Criscione, C.D. An infectious topic in reticulate evolution: introgression and hybridization in animal parasites. *Genes (Basel)* **1**, 102-23 (2010).
194. Cwiklinski, K. *et al.* The *Fasciola hepatica* genome: gene duplication and polymorphism reveals adaptation to the host environment and the capacity for rapid evolution. *Genome Biol* **16**, 71 (2015).
195. Muller, F., Bernard, V. & Tobler, H. Chromatin diminution in nematodes. *Bioessays* **18**, 133-8 (1996).
196. Subirana, J.A. & Messeguer, X. A satellite explosion in the genome of holocentric nematodes. *PLoS One* **8**, e62221 (2013).
197. Streeck, R.E., Moritz, K.B. & Beer, K. Chromatin diminution in *Ascaris suum*: nucleotide sequence of the eliminated satellite DNA. *Nucleic Acids Res* **10**, 3495-502 (1982).
198. Pimpinelli, S. & Goday, C. Unusual kinetochores and chromatin diminution in *Parascaris*. *Trends Genet* **5**, 310-5 (1989).
199. Szitenberg, A. *et al.* Genetic Drift, Not Life History or RNAi, Determine Long-Term Evolution of Transposable Elements. *Genome Biol Evol* **8**, 2964-2978 (2016).
200. van Megen, H. *et al.* A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. *Nematology* **11**, 927-950 (2009).

201. Laetsch, D.R., Heitlinger, E.G., Taraschewski, H., Nadler, S.A. & Blaxter, M.L. The phylogenetics of Anguillicolidae (Nematoda: Anguillicolioidea), swimbladder parasites of eels. *BMC Evol Biol* **12**, 60 (2012).
202. Blaxter, M. & Koutsovoulos, G. The evolution of parasitism in Nematoda. *Parasitology* **142 Suppl 1**, S26-39 (2015).
203. De Ley, P. & Blaxter, M. Systematic Position and Phylogeny. In *The Biology of Nematodes*. (ed. Lee, D.L.) 1-30 (CRC Press, 2002).
204. Park, J.K. *et al.* A common origin of complex life cycles in parasitic flatworms: evidence from the complete mitochondrial genome of *Microcotyle sebastis* (Monogenea: Platyhelminthes). *BMC Evol Biol* **7**, 11 (2007).
205. Lockyer, A.E., Olson, P.D. & Littlewood, D.T.J. Utility of complete large and small subunit rRNA genes in resolving the phylogeny of the Neodermata (Platyhelminthes): implications and a review of the cercomer theory. *Biological Journal of the Linnean Society* **78**, 155-171 (2003).
206. Hahn, C., Fromm, B. & Bachmann, L. Comparative genomics of flatworms (platyhelminthes) reveals shared genomic features of ecto- and endoparasitic neodermata. *Genome Biol Evol* **6**, 1105-17 (2014).
207. Justine, J.L. Non-monophyly of the monogeneans? *Int J Parasitol* **28**, 1653-7 (1998).
208. Olson, P.D. & Littlewood, D.T. Phylogenetics of the Monogenea--evidence from a medley of molecules. *Int J Parasitol* **32**, 233-44 (2002).
209. Jordanova, R. *et al.* Fatty acid- and retinoid-binding proteins have distinct binding pockets for the two types of cargo. *J Biol Chem* **284**, 35818-26 (2009).
210. Finn, R.D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**, D279-85 (2016).
211. O'Donovan, A., Davies, A.A., Moggs, J.G., West, S.C. & Wood, R.D. XPG endonuclease makes the 3' incision in human DNA nucleotide excision repair. *Nature* **371**, 432-5 (1994).
212. Habraken, Y., Sung, P., Prakash, L. & Prakash, S. Yeast excision repair gene RAD2 encodes a single-stranded DNA endonuclease. *Nature* **366**, 365-8 (1993).
213. Eyboullet, F. *et al.* Mediator links transcription and DNA repair by facilitating Rad2/XPG recruitment. *Genes Dev* **27**, 2549-62 (2013).
214. Ruelas, D.S., Karentz, D. & Sullivan, J.T. Sublethal effects of ultraviolet b radiation on miracidia and sporocysts of *Schistosoma mansoni*: intramolluscan development, infectivity, and photoreactivation. *J Parasitol* **93**, 1303-10 (2007).
215. Silva, C.S. *et al.* *Schistosoma mansoni*: gene expression of the nucleotide excision repair factor 2 (NEF2) during the parasite life cycle, and in adult worms after exposure to different DNA-damaging agents. *Acta Trop* **104**, 52-62 (2007).
216. Studer, A., Lamare, M.D. & Poulin, R. Effects of ultraviolet radiation on the transmission process of an intertidal trematode parasite. *Parasitology* **139**, 537-46 (2012).
217. Davies, K.G. Understanding the interaction between an obligate hyperparasitic bacterium, *Pasteuria penetrans* and its obligate plant-parasitic nematode host, *Meloidogyne* spp. *Adv Parasitol* **68**, 211-45 (2009).
218. Alper, S., McBride, S.J., Lackford, B., Freedman, J.H. & Schwartz, D.A. Specificity and complexity of the *Caenorhabditis elegans* innate immune response. *Mol Cell Biol* **27**, 5544-53 (2007).
219. Schulenburg, H. & Boehnisch, C. Diversification and adaptive sequence evolution of *Caenorhabditis* lysozymes (Nematoda: Rhabditidae). *BMC Evol Biol* **8**, 114 (2008).
220. Marsh, E.K., van den Berg, M.C. & May, R.C. A two-gene balance regulates *Salmonella typhimurium* tolerance in the nematode *Caenorhabditis elegans*. *PLoS One* **6**, e16839 (2011).

221. Boehnisch, C. *et al.* Protist-type lysozymes of the nematode *Caenorhabditis elegans* contribute to resistance against pathogenic *Bacillus thuringiensis*. *PLoS One* **6**, e24619 (2011).
222. Mulvenna, J. *et al.* Proteomics analysis of the excretory/secretory component of the blood-feeding stage of the hookworm, *Ancylostoma caninum*. *Mol Cell Proteomics* **8**, 109-21 (2009).
223. Chavez, V., Mohri-Shiomi, A. & Garsin, D.A. Ce-Duox1/BLI-3 generates reactive oxygen species as a protective innate immune mechanism in *Caenorhabditis elegans*. *Infect Immun* **77**, 4983-9 (2009).
224. Walton, A.C. The oogenesis and early embryology of *Ascaris canis*. *Journal of Morphology* **30**, 527-603 (1918).
225. Graser, S., Stierhof, Y.D. & Nigg, E.A. Cep68 and Cep215 (Cdk5rap2) are required for centrosome cohesion. *J Cell Sci* **120**, 4321-31 (2007).
226. Fisk Green, R., Lorson, M., Walhout, A.J., Vidal, M. & van den Heuvel, S. Identification of critical domains and putative partners for the *Caenorhabditis elegans* spindle component LIN-5. *Mol Genet Genomics* **271**, 532-44 (2004).
227. Perry, R.N. & Wharton, D.A. *Molecular and Physiological Basis of Nematode Survival*, (CABI, 2011).
228. Sonobe, H. & Ito, Y. Phosphoconjugation and dephosphorylation reactions of steroid hormone in insects. *Mol Cell Endocrinol* **307**, 25-35 (2009).
229. Sonobe, H. *et al.* Purification, kinetic characterization, and molecular cloning of a novel enzyme, ecdysteroid 22-kinase. *J Biol Chem* **281**, 29513-24 (2006).
230. Sluder, A.E. & Maina, C.V. Nuclear receptors in nematodes: themes and variations. *Trends Genet* **17**, 206-13 (2001).
231. Tzertzinis, G. *et al.* Molecular evidence for a functional ecdysone signaling system in *Brugia malayi*. *PLoS Negl Trop Dis* **4**, e625 (2010).
232. Parihar, M. *et al.* The genome of the nematode *Pristionchus pacificus* encodes putative homologs of RXR/Usp and EcR. *Gen Comp Endocrinol* **167**, 11-7 (2010).
233. Shea, C., Richer, J., Tzertzinis, G. & Maina, C.V. An EcR homolog from the filarial parasite, *Dirofilaria immitis* requires a ligand-activated partner for transactivation. *Mol Biochem Parasitol* **171**, 55-63 (2010).
234. Graham, L.D., Kotze, A.C., Fernley, R.T. & Hill, R.J. An ortholog of the ecdysone receptor protein (EcR) from the parasitic nematode *Haemonchus contortus*. *Mol Biochem Parasitol* **171**, 104-7 (2010).
235. Zhao, L. *et al.* Chemical signals synchronize the life cycles of a plant-parasitic nematode and its vector beetle. *Curr Biol* **23**, 2038-43 (2013).
236. Collins, J.J., 3rd. Platyhelminthes. *Curr Biol* **27**, R252-R256 (2017).
237. Uccelletti, D. *et al.* APY-1, a novel *Caenorhabditis elegans* apyrase involved in unfolded protein response signalling and stress responses. *Mol Biol Cell* **19**, 1337-45 (2008).
238. Valenzuela, J.G., Charlab, R., Galperin, M.Y. & Ribeiro, J.M. Purification, cloning, and expression of an apyrase from the bed bug *Cimex lectularius*. A new type of nucleotide-binding enzyme. *J Biol Chem* **273**, 30583-90 (1998).
239. Nisbet, A.J. *et al.* A calcium-activated apyrase from *Teladorsagia circumcincta*: an excretory/secretory antigen capable of modulating host immune responses? *Parasite Immunol* **33**, 236-43 (2011).
240. Hewitson, J.P. *et al.* Secretion of protective antigens by tissue-stage nematode larvae revealed by proteomic analysis and vaccination-induced sterile immunity. *PLoS Pathog* **9**, e1003492 (2013).

241. Tsugeno, Y. & Ito, A. A key amino acid responsible for substrate selectivity of monoamine oxidase A and B. *J Biol Chem* **272**, 14033-6 (1997).
242. Smolinska, S., Jutel, M., Cramer, R. & O'Mahony, L. Histamine and gut mucosal immune regulation. *Allergy* **69**, 273-81 (2014).
243. Guevara-Flores, A. *et al.* 5'-p-Fluorosulfonyl benzoyl adenosine inhibits an ecto-ATP-diphosphohydrolase in the tegument surface of *Taenia crassiceps* cysticerci. *Mol Biochem Parasitol* **162**, 123-33 (2008).
244. Wu, C. *et al.* Non-immune immunoglobulins shield *Schistosoma japonicum* from host immunorecognition. *Sci Rep* **5**, 13434 (2015).
245. Courtney, H.S. & Li, Y. Non-immune binding of human IgG to M-related proteins confers resistance to phagocytosis of group A streptococci in blood. *PLoS One* **8**, e78719 (2013).
246. Silva, L.L. *et al.* The *Schistosoma mansoni* phylome: using evolutionary genomics to gain insight into a parasite's biology. *BMC Genomics* **13**, 617 (2012).
247. Cabral, G.A. Lipids as bioeffectors in the immune system. *Life Sci* **77**, 1699-710 (2005).
248. Basavaraju, S.V. *et al.* Ac-FAR-1, a 20 kDa fatty acid- and retinol-binding protein secreted by adult *Ancylostoma caninum* hookworms: gene transcription pattern, ligand binding properties and structural characterisation. *Mol Biochem Parasitol* **126**, 63-71 (2003).
249. Garofalo, A. *et al.* The FAR proteins of filarial nematodes: secretion, glycosylation and lipid binding characteristics. *Mol Biochem Parasitol* **122**, 161-70 (2002).
250. Schwarz, E.M. *et al.* The genome and developmental transcriptome of the strongylid nematode *Haemonchus contortus*. *Genome Biol* **14**, R89 (2013).
251. Vanholme, B. *et al.* Detection of putative secreted proteins in the plant-parasitic nematode *Heterodera schachtii*. *Parasitol Res* **98**, 414-24 (2006).
252. Bosman, F.T. & Stamenkovic, I. Functional structure and composition of the extracellular matrix. *J Pathol* **200**, 423-8 (2003).
253. Varki, A. Biological roles of glycans. *Glycobiology* **27**, 3-49 (2017).
254. Varki, A. & Lowe, J.B. Biological Roles of Glycans. In *Essentials of Glycobiology* (eds. Varki, A. *et al.*) (Cold Spring Harbor (NY), 2009).
255. Kaneiwa, T. *et al.* Identification of a novel chondroitin hydrolase in *Caenorhabditis elegans*. *J Biol Chem* **283**, 14971-9 (2008).
256. Irwin, J.A. *et al.* Glycosidase activity in the excretory-secretory products of the liver fluke, *Fasciola hepatica*. *Parasitology* **129**, 465-72 (2004).
257. Bruschi, F. & Chiumiento, L. *Trichinella* inflammatory myopathy: host or parasite strategy? *Parasit Vectors* **4**, 42 (2011).
258. Tilney, L.G., Connelly, P.S., Guild, G.M., Vranich, K.A. & Artis, D. Adaptation of a nematode parasite to living within the mammalian epithelium. *J Exp Zool A Comp Exp Biol* **303**, 927-45 (2005).
259. Marchant, J. *et al.* Galactose recognition by the apicomplexan parasite *Toxoplasma gondii*. *J Biol Chem* **287**, 16720-33 (2012).
260. Diaz, A. *et al.* Understanding the laminated layer of larval *Echinococcus* I: structure. *Trends Parasitol* **27**, 204-13 (2011).
261. Arana, Y. *et al.* Characterization of the carbohydrate components of *Taenia solium* oncosphere proteins and their role in the antigenicity. *Parasitol Res* **112**, 3569-78 (2013).
262. Maruyama, H., El-Malky, M., Kumagai, T. & Ohta, N. Secreted adhesion molecules of *Strongyloides venezuelensis* are produced by oesophageal glands and are components of the wall of tunnels constructed by adult worms in the host intestinal mucosa. *Parasitology* **126**, 165-71 (2003).
263. Fairbairn, D. Biochemical adaptation and loss of genetic capacity in helminth parasites. *Biol Rev Camb Philos Soc* **45**, 29-72 (1970).

264. Murray, P., Rosenthal, K. & Pfaller, M. *Medical Microbiology* (Mosby, 2005).
265. Chalmers, I.W. & Hoffmann, K.F. Platyhelminth Venom Allergen-Like (VAL) proteins: revealing structural diversity, class-specific features and biological associations across the phylum. *Parasitology* **139**, 1231-45 (2012).
266. Hawdon, J.M., Jones, B.F., Hoffman, D.R. & Hotez, P.J. Cloning and characterization of Ancylostoma-secreted protein. A novel protein associated with the transition to parasitism by infective hookworm larvae. *J Biol Chem* **271**, 6672-8 (1996).
267. Hawdon, J.M., Narasimhan, S. & Hotez, P.J. Ancylostoma secreted protein 2: cloning and characterization of a second member of a family of nematode secreted proteins from Ancylostoma caninum. *Mol Biochem Parasitol* **99**, 149-65 (1999).
268. Cantacessi, C. *et al.* Insights into SCP/TAPS proteins of liver flukes based on large-scale bioinformatic analyses of sequence datasets. *PLoS One* **7**, e31164 (2012).
269. Curwen, R.S., Ashton, P.D., Sundaralingam, S. & Wilson, R.A. Identification of novel proteases and immunomodulators in the secretions of schistosome cercariae that facilitate host entry. *Mol Cell Proteomics* **5**, 835-44 (2006).
270. Bower, M.A., Constant, S.L. & Mendez, S. Necator americanus: the Na-ASP-2 protein secreted by the infective larvae induces neutrophil recruitment in vivo and in vitro. *Exp Parasitol* **118**, 569-75 (2008).
271. Bethony, J. *et al.* Antibodies against a secreted protein from hookworm larvae reduce the intensity of hookworm infection in humans and vaccinated laboratory animals. *FASEB J* **19**, 1743-5 (2005).
272. Mendez, S., A, D.S., Antoine, A.D., Ahn, S. & Hotez, P. Use of the air pouch model to investigate immune responses to a hookworm vaccine containing the Na-ASP-2 protein in rats. *Parasite Immunol* **30**, 53-6 (2008).
273. Xiao, S. *et al.* The evaluation of recombinant hookworm antigens as vaccines in hamsters (Mesocricetus auratus) challenged with human hookworm, Necator americanus. *Exp Parasitol* **118**, 32-40 (2008).
274. Darwiche, R. *et al.* Crystal structure of Brugia malayi venom allergen-like protein-1 (BmVAL-1), a vaccine candidate for lymphatic filariasis. *Int J Parasitol* **48**, 371-378 (2018).
275. Schwarz, E.M. *et al.* The genome and transcriptome of the zoonotic hookworm Ancylostoma ceylanicum identify infection-specific gene families. *Nat Genet* **47**, 416-22 (2015).
276. Robellet, X., Flippin, M., Pegot, S., Maccabe, A.P. & Velot, C. AcpA, a member of the GPR1/FUN34/YaaH membrane protein family, is essential for acetate permease activity in the hyphal fungus Aspergillus nidulans. *Biochem J* **412**, 485-93 (2008).
277. Sa-Pessoa, J. *et al.* SATP (YaaH), a succinate-acetate transporter protein in Escherichia coli. *Biochem J* **454**, 585-95 (2013).
278. Goodenough, U. *et al.* The path to triacylglyceride obesity in the sta6 strain of Chlamydomonas reinhardtii. *Eukaryot Cell* **13**, 591-613 (2014).
279. Kmetec, E. & Bueding, E. Production of succinate by the canine whipworm Trichuris vulpis. *Comp Biochem Physiol* **15**, 271-4 (1965).
280. Palkova, Z. *et al.* Ammonia pulses and metabolic oscillations guide yeast colony development. *Mol Biol Cell* **13**, 3901-14 (2002).
281. Bader, G. *et al.* Crystal structure of rat GTP cyclohydrolase I feedback regulatory protein, GFRP. *J Mol Biol* **312**, 1051-7 (2001).
282. Loer, C.M. *et al.* Cuticle integrity and biogenic amine synthesis in Caenorhabditis elegans require the cofactor tetrahydrobiopterin (BH4). *Genetics* **200**, 237-53 (2015).

283. Funderburk, C.D., Bowling, K.M., Xu, D., Huang, Z. & O'Donnell, J.M. A typical N-terminal extensions confer novel regulatory properties on GTP cyclohydrolase isoforms in *Drosophila melanogaster*. *J Biol Chem* **281**, 33302-12 (2006).
284. Soblik, H. *et al.* Life cycle stage-resolved proteomic analysis of the excretome/secretome from *Strongyloides ratti*--identification of stage-specific proteases. *Mol Cell Proteomics* **10**, M111 010157 (2011).
285. Feng, J. *et al.* Molecular cloning and characterization of Ac-MTP-2, an astacin-like metalloprotease released by adult *Ancylostoma caninum*. *Mol Biochem Parasitol* **152**, 132-8 (2007).
286. Gomez Gallego, S. *et al.* Identification of an astacin-like metallo-proteinase transcript from the infective larvae of *Strongyloides stercoralis*. *Parasitol Int* **54**, 123-33 (2005).
287. Dvorak, J. *et al.* Differential use of protease families for invasion by schistosome cercariae. *Biochimie* **90**, 345-58 (2008).
288. Robinson, M.W., Dalton, J.P. & Donnelly, S. Helminth pathogen cathepsin proteases: it's a family affair. *Trends Biochem Sci* **33**, 601-8 (2008).
289. Wilson, L.R. *et al.* *Fasciola hepatica*: characterization and cloning of the major cathepsin B protease secreted by newly excysted juvenile liver fluke. *Exp Parasitol* **88**, 85-94 (1998).
290. Pratt, D., Cox, G.N., Milhausen, M.J. & Boisvenue, R.J. A developmentally regulated cysteine protease gene family in *Haemonchus contortus*. *Mol Biochem Parasitol* **43**, 181-91 (1990).
291. Pratt, D. *et al.* Cloning and sequence comparisons of four distinct cysteine proteases expressed by *Haemonchus contortus* adult worms. *Mol Biochem Parasitol* **51**, 209-18 (1992).
292. Klinkert, M.Q., Felleisen, R., Link, G., Ruppel, A. & Beck, E. Primary structures of Sm31/32 diagnostic proteins of *Schistosoma mansoni* and their identification as proteases. *Mol Biochem Parasitol* **33**, 113-22 (1989).
293. Rehman, A. & Jasmer, D.P. A tissue specific approach for analysis of membrane and secreted protein antigens from *Haemonchus contortus* gut and its application to diverse nematode species. *Mol Biochem Parasitol* **97**, 55-68 (1998).
294. Skuce, P.J. *et al.* Molecular cloning and characterization of gut-derived cysteine proteinases associated with a host protective extract from *Haemonchus contortus*. *Parasitology* **119** (Pt 4), 405-12 (1999).
295. Harrop, S.A., Sawangjaroen, N., Prociv, P. & Brindley, P.J. Characterization and localization of cathepsin B proteinases expressed by adult *Ancylostoma caninum* hookworms. *Mol Biochem Parasitol* **71**, 163-71 (1995).
296. Cantacessi, C. *et al.* Differences in transcription between free-living and CO₂-activated third-stage larvae of *Haemonchus contortus*. *BMC Genomics* **11**, 266 (2010).
297. Williamson, A.L., Brindley, P.J., Knox, D.P., Hotez, P.J. & Loukas, A. Digestive proteases of blood-feeding nematodes. *Trends Parasitol* **19**, 417-23 (2003).
298. Krautz-Peterson, G. & Skelly, P.J. Schistosome asparaginyl endopeptidase (legumain) is not essential for cathepsin B1 activation in vivo. *Mol Biochem Parasitol* **159**, 54-8 (2008).
299. Beckham, S.A. *et al.* A major cathepsin B protease from the liver fluke *Fasciola hepatica* has atypical active site features and a potential role in the digestive tract of newly excysted juvenile parasites. *Int J Biochem Cell Biol* **41**, 1601-12 (2009).
300. Cancela, M. *et al.* Survey of transcripts expressed by the invasive juvenile stage of the liver fluke *Fasciola hepatica*. *BMC Genomics* **11**, 227 (2010).
301. McGonigle, L. *et al.* The silencing of cysteine proteases in *Fasciola hepatica* newly excysted juveniles using RNA interference reduces gut penetration. *Int J Parasitol* **38**, 149-55 (2008).

302. Liu, R.D. *et al.* Screening and characterization of early diagnostic antigens in excretory-secretory proteins from *Trichinella spiralis* intestinal infective larvae by immunoproteomics. *Parasitol Res* **115**, 615-22 (2016).
303. Trap, C. *et al.* Cloning and analysis of a cDNA encoding a putative serine protease comprising two trypsin-like domains of *Trichinella spiralis*. *Parasitol Res* **98**, 288-94 (2006).
304. Young, N.D. *et al.* The *Opisthorchis viverrini* genome provides insights into life in the bile duct. *Nat Commun* **5**, 4378 (2014).
305. de Boer, J.P. *et al.* Alpha-2-macroglobulin functions as an inhibitor of fibrinolytic, clotting, and neutrophilic proteinases in sepsis: studies using a baboon model. *Infect Immun* **61**, 5035-43 (1993).
306. Sripa, B. *et al.* The tumorigenic liver fluke *Opisthorchis viverrini*--multiple pathways to cancer. *Trends Parasitol* **28**, 395-407 (2012).
307. Kim, T.H., Kim, Y.J., Cho, J.W. & Shim, J. A novel zinc-carboxypeptidase SURO-1 regulates cuticle formation and body morphogenesis in *Caenorhabditis elegans*. *FEBS Lett* **585**, 121-7 (2011).
308. Brennan, R.G. & Matthews, B.W. The helix-turn-helix DNA binding motif. *J Biol Chem* **264**, 1903-6 (1989).
309. Chen, N. *et al.* Identification of a nematode chemosensory gene family. *Proc Natl Acad Sci U S A* **102**, 146-51 (2005).
310. Thomas, J.H., Kelley, J.L., Robertson, H.M., Ly, K. & Swanson, W.J. Adaptive evolution in the SRZ chemoreceptor families of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Proc Natl Acad Sci U S A* **102**, 4476-81 (2005).
311. Park, D. *et al.* Interaction of structure-specific and promiscuous G-protein-coupled receptors mediates small-molecule signaling in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A* **109**, 9917-22 (2012).
312. Zamanian, M. *et al.* The repertoire of G protein-coupled receptors in the human parasite *Schistosoma mansoni* and the model organism *Schmidtea mediterranea*. *BMC Genomics* **12**, 596 (2011).
313. Cully, D.F. & Pares, P.S. Solubilization and characterization of a high affinity ivermectin binding site from *Caenorhabditis elegans*. *Mol Pharmacol* **40**, 326-32 (1991).
314. Jones, A.K. & Sattelle, D.B. The cys-loop ligand-gated ion channel gene superfamily of the nematode, *Caenorhabditis elegans*. *Invert Neurosci* **8**, 41-7 (2008).
315. Dufour, V., Beech, R.N., Wever, C., Dent, J.A. & Geary, T.G. Molecular cloning and characterization of novel glutamate-gated chloride channel subunits from *Schistosoma mansoni*. *PLoS Pathog* **9**, e1003586 (2013).
316. Beg, A.A. & Jorgensen, E.M. EXP-1 is an excitatory GABA-gated cation channel. *Nat Neurosci* **6**, 1145-52 (2003).
317. Dent, J.A. Evidence for a diverse Cys-loop ligand-gated ion channel superfamily in early bilateria. *J Mol Evol* **62**, 523-35 (2006).
318. Putrenko, I., Zakikhani, M. & Dent, J.A. A family of acetylcholine-gated chloride channel subunits in *Caenorhabditis elegans*. *J Biol Chem* **280**, 6392-8 (2005).
319. Ribeiro, P., Gupta, V. & El-Sakkary, N. Biogenic amines and the control of neuromuscular signaling in schistosomes. *Invert Neurosci* **12**, 13-28 (2012).
320. Theodoulou, F.L. & Kerr, I.D. ABC transporter research: going strong 40 years on. *Biochem Soc Trans* **43**, 1033-40 (2015).
321. Sheps, J.A., Ralph, S., Zhao, Z., Baillie, D.L. & Ling, V. The ABC transporter gene family of *Caenorhabditis elegans* has implications for the evolutionary dynamics of multidrug resistance in eukaryotes. *Genome Biol* **5**, R15 (2004).

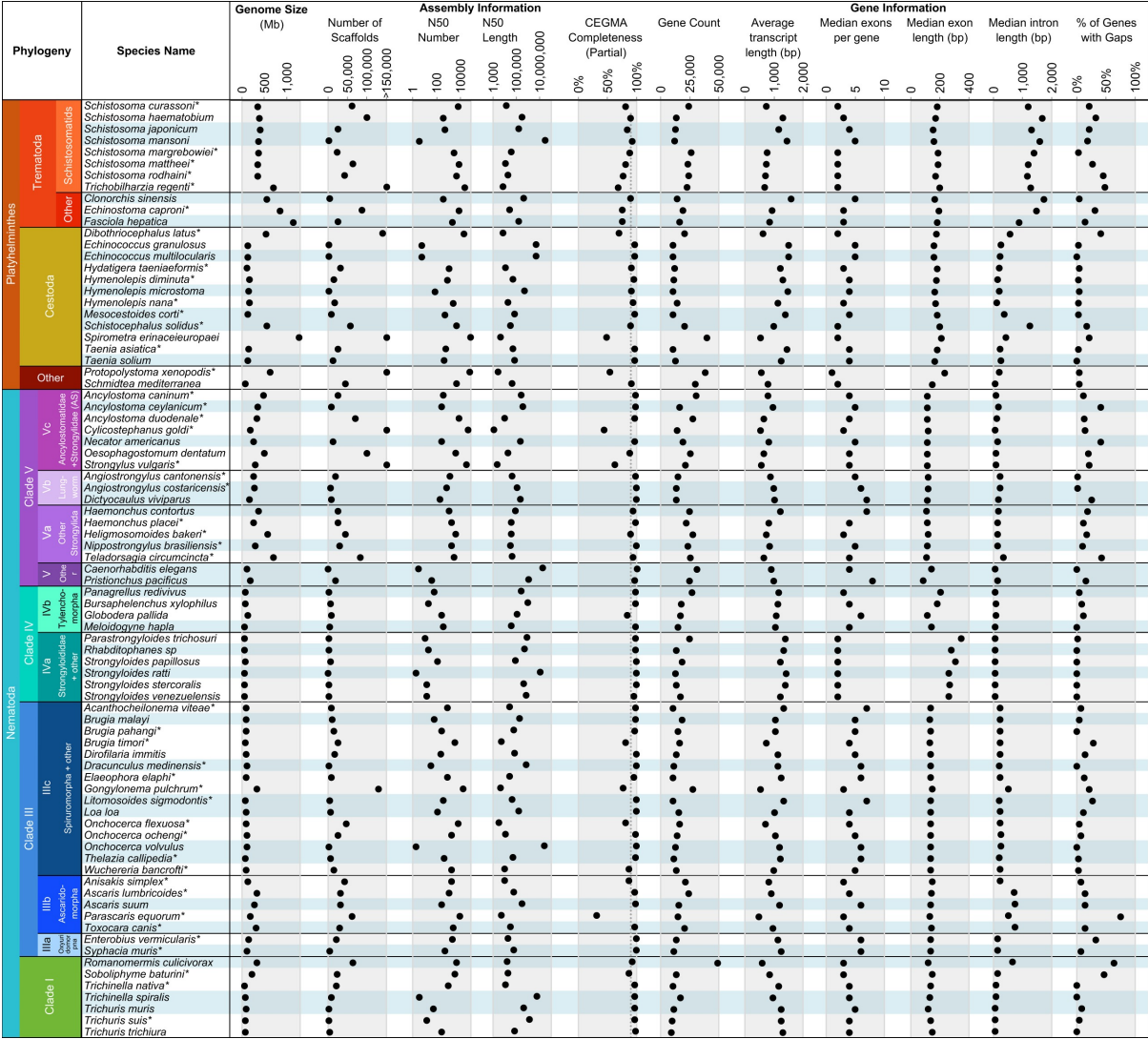
322. Schumacher, T. & Benndorf, R.A. ABC Transport Proteins in Cardiovascular Disease-A Brief Summary. *Molecules* **22**(2017).
323. Lincke, C.R., Broeks, A., The, I., Plasterk, R.H. & Borst, P. The expression of two P-glycoprotein (pgp) genes in transgenic *Caenorhabditis elegans* is confined to intestinal cells. *EMBO J* **12**, 1615-20 (1993).
324. Kurz, C.L., Shapira, M., Chen, K., Baillie, D.L. & Tan, M.W. *Caenorhabditis elegans* pgp-5 is involved in resistance to bacterial infection and heavy metal and its regulation requires TIR-1 and a p38 map kinase cascade. *Biochem Biophys Res Commun* **363**, 438-43 (2007).
325. Issouf, M. *et al.* *Haemonchus contortus* P-glycoproteins interact with host eosinophil granules: a novel insight into the role of ABC transporters in host-parasite interaction. *PLoS One* **9**, e87802 (2014).
326. Kerboeuf, D., Blackhall, W., Kaminsky, R. & von Samson-Himmelstjerna, G. P-glycoprotein in helminths: function and perspectives for anthelmintic treatment and reversal of resistance. *Int J Antimicrob Agents* **22**, 332-46 (2003).
327. Bosch, I.B., Wang, Z.X., Tao, L.F. & Shoemaker, C.B. Two *Schistosoma mansoni* cDNAs encoding ATP-binding cassette (ABC) family proteins. *Mol Biochem Parasitol* **65**, 351-6 (1994).
328. Mapes, J. *et al.* CED-1, CED-7, and TTR-52 regulate surface phosphatidylserine expression on apoptotic and phagocytic cells. *Curr Biol* **22**, 1267-75 (2012).
329. Furlong, S.T. & Caulfield, J.P. *Schistosoma mansoni*: sterol and phospholipid composition of cercariae, schistosomula, and adults. *Exp Parasitol* **65**, 222-31 (1988).
330. Chen, D. *et al.* Germline signaling mediates the synergistically prolonged longevity produced by double mutations in *daf-2* and *rsk-1* in *C. elegans*. *Cell Rep* **5**, 1600-10 (2013).
331. Hashiguchi, M. & Hashiguchi, T. Kinase-kinase interaction and modulation of tau phosphorylation. *Int Rev Cell Mol Biol* **300**, 121-60 (2013).
332. Liao, J.C., Yang, T.T., Weng, R.R., Kuo, C.T. & Chang, C.W. TTBK2: a tau protein kinase beyond tau phosphorylation. *Biomed Res Int* **2015**, 575170 (2015).
333. LaMunyon, C.W. *et al.* A New Player in the Spermiogenesis Pathway of *Caenorhabditis elegans*. *Genetics* **201**, 1103-16 (2015).
334. Muhlrads, P.J., Clark, J.N., Nasri, U., Sullivan, N.G. & LaMunyon, C.W. SPE-8, a protein-tyrosine kinase, localizes to the spermatid cell membrane through interaction with other members of the SPE-8 group spermatid activation signaling pathway in *C. elegans*. *BMC Genet* **15**, 83 (2014).
335. Mila, D. *et al.* Asymmetric Wnt Pathway Signaling Facilitates Stem Cell-Like Divisions via the Nonreceptor Tyrosine Kinase FRK-1 in *Caenorhabditis elegans*. *Genetics* **201**, 1047-60 (2015).
336. Korting, W. & Fairbairn, D. Changes in beta-oxidation and related enzymes during the life cycle of *Strongyloides ratti* (Nematoda). *J Parasitol* **57**, 1153-8 (1971).
337. Barrett, J. & Korting, W. Lipid Catabolism in Plerocercoids of *Schistocephalus Solidus* (Cestoda-Pseudophyllidea). *International Journal for Parasitology* **7**, 419-422 (1977).
338. Barrett, J. & Korting, W. Studies on beta-oxidation in the adult liver fluke *Fasciola hepatica*. *Int J Parasitol* **6**, 155-7 (1976).
339. 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**, 1009-12 (1970).
340. Fraga, C.M. *et al.* Alternative energy production pathways in *Taenia crassiceps* cysticerci in vitro exposed to a benzimidazole derivative (RCB20). *Parasitology* **143**, 488-93 (2016).

341. Skelly, P.J., Da'dara, A.A., Li, X.H., Castro-Borges, W. & Wilson, R.A. Schistosome feeding and regurgitation. *PLoS Pathog* **10**, e1004246 (2014).
342. Rodriguez-Contreras, D., Skelly, P.J., Landa, A., Shoemaker, C.B. & Laclette, J.P. Molecular and functional characterization and tissue localization of 2 glucose transporter homologues (TGTP1 and TGTP2) from the tapeworm *Taenia solium*. *Parasitology* **117** (Pt 6), 579-88 (1998).
343. Barrett, J., Ward, C.W. & Fairbairn, D. Glyoxylate Cycle and Conversion of Triglycerides to Carbohydrates in Developing Eggs of *Ascaris-Lumbricoides*. *Comparative Biochemistry and Physiology* **35**, 577-586 (1970).
344. Komuniecki, R. & Harris, B.G. Carbohydrate and Energy Metabolism in Helminths. In *Biochemistry and Molecular Biology of Parasites*. (eds. Marr, J.J. & Muller, M.) 49-66 (Academic Press, 1995).
345. Madin, K.A.C., Loomis, S.H. & Crowe, J.H. Anhydrobiosis in Nematodes - Control of Carbon Flow through the Glyoxylate Cycle. *Journal of Experimental Zoology* **234**, 341-350 (1985).
346. Barrett, J. Forty years of helminth biochemistry. *Parasitology* **136**, 1633-42 (2009).
347. Prichard, R.K. & Schofield, P.J. The glyoxylate cycle, fructose-1,6-diphosphatase and glyconeogenesis in *Fasciola hepatica*. *Comparative Biochemistry and Physiology* **29**, 581-590 (1968).
348. Rafi, M.M. & Raj, R.K. Phosphoenolpyruvate-Succinate-Glyoxylate Pathway in the Filarial Parasite *Setaria-Digitata*. *Journal of Biosciences* **16**, 121-126 (1991).
349. Voet, D. & Voet, J.G. *Biochemistry*, (John Wiley & Sons, 2011).
350. Okubo, Y., Yang, S., Chistoserdova, L. & Lidstrom, M.E. Alternative route for glyoxylate consumption during growth on two-carbon compounds by *Methylobacterium extorquens* AM1. *J Bacteriol* **192**, 1813-23 (2010).
351. Wang, T. *et al.* Proteomic analysis of the excretory-secretory products from larval stages of *Ascaris suum* reveals high abundance of glycosyl hydrolases. *PLoS Negl Trop Dis* **7**, e2467 (2013).
352. Braeckman, B.P., Houthoofd, K. & Vanfleteren, J.R. Intermediary metabolism. *WormBook*, 1-24 (2009).
353. Muller, M. *et al.* Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol Mol Biol Rev* **76**, 444-95 (2012).
354. Maule, A.G. & Marks, N.J. (eds.). *Parasitic flatworms: molecular biology, biochemistry, immunology and physiology*, (CABI, 2005).
355. Kohler, P. The strategies of energy conservation in helminths. *Mol Biochem Parasitol* **17**, 1-18 (1985).
356. Inaoka, D.K. *et al.* Structural Insights into the Molecular Design of Flutolanil Derivatives Targeted for Fumarate Respiration of Parasite Mitochondria. *Int J Mol Sci* **16**, 15287-308 (2015).
357. Kim, J., Hannibal, L., Gherasim, C., Jacobsen, D.W. & Banerjee, R. A human vitamin B12 trafficking protein uses glutathione transferase activity for processing alkylcobalamins. *J Biol Chem* **284**, 33418-24 (2009).
358. Froese, D.S. *et al.* Structural Insights into the MMACHC-MMADHC Protein Complex Involved in Vitamin B12 Trafficking. *J Biol Chem* **290**, 29167-77 (2015).
359. Yamada, K., Gherasim, C., Banerjee, R. & Koutmos, M. Structure of Human B12 Trafficking Protein CblD Reveals Molecular Mimicry and Identifies a New Subfamily of Nitro-FMN Reductases. *J Biol Chem* **290**, 29155-66 (2015).
360. Barrett, J. *Biochemistry of parasitic helminths*, (MacMillan Publishers Ltd., London, 1981).

361. Roth, J.R., Lawrence, J.G. & Bobik, T.A. Cobalamin (coenzyme B12): synthesis and biological significance. *Annu Rev Microbiol* **50**, 137-81 (1996).
362. Glenwright, A.J. *et al.* Structural basis for nutrient acquisition by dominant members of the human gut microbiota. *Nature* **541**, 407-411 (2017).
363. Woodson, J.D., Zayas, C.L. & Escalante-Semerena, J.C. A new pathway for salvaging the coenzyme B12 precursor cobinamide in archaea requires cobinamide-phosphate synthase (CbiB) enzyme activity. *J Bacteriol* **185**, 7193-201 (2003).
364. Rosa, B.A., Jasmer, D.P. & Mitreva, M. Genome-wide tissue-specific gene expression, co-expression and regulation of co-expressed genes in adult nematode *Ascaris suum*. *PLoS Negl Trop Dis* **8**, e2678 (2014).
365. Tielens, A.G., van Grinsven, K.W., Henze, K., van Hellemond, J.J. & Martin, W. Acetate formation in the energy metabolism of parasitic helminths and protists. *Int J Parasitol* **40**, 387-97 (2010).
366. Von Brand, T. *Biochemistry of parasites*, (Academic Press, 1973).
367. Jin, L. *et al.* Glutamate dehydrogenase 1 signals through antioxidant glutathione peroxidase 1 to regulate redox homeostasis and tumor growth. *Cancer Cell* **27**, 257-70 (2015).
368. Mustafa, T., Komuniecki, R. & Mettrick, D.F. Cytosolic Glutamate-Dehydrogenase in Adult *Hymenolepis Diminuta* (Cestoda). *Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology* **61**, 219-222 (1978).
369. Skuce, P.J., Stewart, E.M., Smith, W.D. & Knox, D.P. Cloning and characterization of glutamate dehydrogenase (GDH) from the gut of *Haemonchus contortus*. *Parasitology* **118** (Pt 3), 297-304 (1999).
370. Kikuchi, G., Motokawa, Y., Yoshida, T. & Hiraga, K. Glycine cleavage system: reaction mechanism, physiological significance, and hyperglycinemia. *Proc Jpn Acad Ser B Phys Biol Sci* **84**, 246-63 (2008).
371. Zheng, H. *et al.* The genome of the hydatid tapeworm *Echinococcus granulosus*. *Nat Genet* **45**, 1168-75 (2013).
372. Mukherjee, M., Brown, M.T., McArthur, A.G. & Johnson, P.J. Proteins of the glycine decarboxylase complex in the hydrogenosome of *Trichomonas vaginalis*. *Eukaryot Cell* **5**, 2062-71 (2006).
373. Jerlstrom-Hultqvist, J. *et al.* Hydrogenosomes in the diplomonad *Spironucleus salmonicida*. *Nat Commun* **4**, 2493 (2013).
374. Payne, S.H. & Loomis, W.F. Retention and loss of amino acid biosynthetic pathways based on analysis of whole-genome sequences. *Eukaryot Cell* **5**, 272-6 (2006).
375. Barrett, J. Amino acid metabolism in helminths. *Adv Parasitol* **30**, 39-105 (1991).
376. Kurelec, B. Aspartate transcarbamylase in some parasitic platyhelminths. *Comp Biochem Physiol B* **47**, 33-40 (1974).
377. Hill, B., Kilsby, J., Rogerson, G.W., McIntosh, R.T. & Ginger, C.D. The enzymes of pyrimidine biosynthesis in a range of parasitic protozoa and helminths. *Mol Biochem Parasitol* **2**, 123-34 (1981).
378. Heath, R.L. & Hart, J.L. Biosynthesis de novo of purines and pyrimidines in *Mesocostoides* (Cestoda). II. *J Parasitol* **56**, 340-5 (1970).
379. Foster, J. *et al.* The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS Biol* **3**, e121 (2005).
380. Jaffe, J.J. & Chrin, L.R. Involvement of tetrahydrofolate cofactors in de novo purine ribonucleotide synthesis by adult *Brugia pahangi* and *Dirofilaria immitis*. *Mol Biochem Parasitol* **2**, 259-70 (1981).

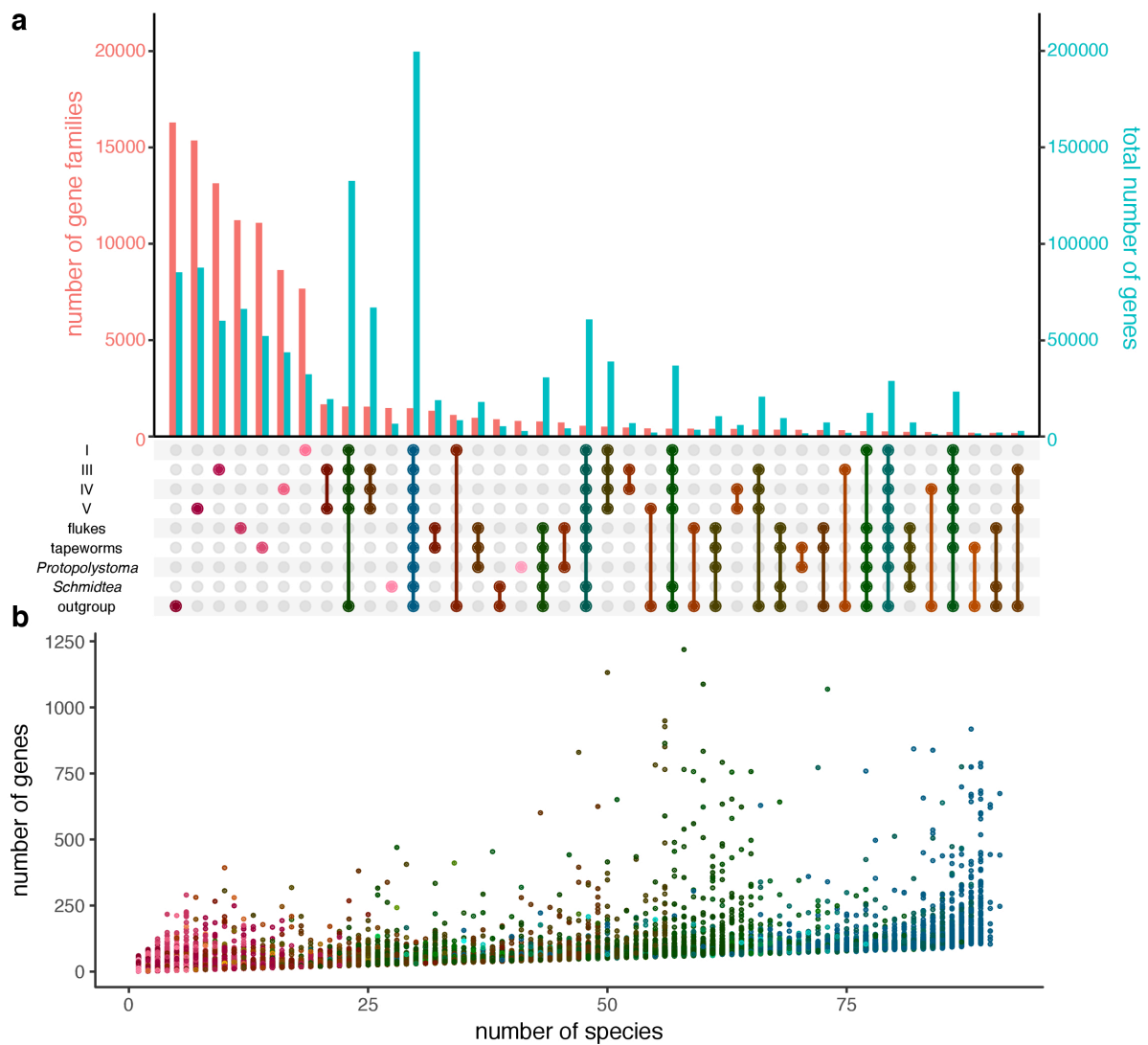
381. Toh, S.Q., Glanfield, A., Gobert, G.N. & Jones, M.K. Heme and blood-feeding parasites: friends or foes? *Parasit Vectors* **3**, 108 (2010).
382. Luck, A.N. *et al.* Heme acquisition in the parasitic filarial nematode *Brugia malayi*. *FASEB J* **30**, 3501-3514 (2016).
383. Glanfield, A., McManus, D.P., Anderson, G.J. & Jones, M.K. Pumping iron: a potential target for novel therapeutics against schistosomes. *Trends Parasitol* **23**, 583-8 (2007).
384. Stubenhaus, B.M. *et al.* Light-induced depigmentation in planarians models the pathophysiology of acute porphyrias. *Elife* **5**(2016).
385. de Jong, L., Meng, Y., Dent, J. & Hekimi, S. Thiamine pyrophosphate biosynthesis and transport in the nematode *Caenorhabditis elegans*. *Genetics* **168**, 845-54 (2004).
386. Walker, J. & Barrett, J. Pyridoxal 5'-phosphate dependent enzymes in the nematode *Nippostrongylus brasiliensis*. *Int J Parasitol* **21**, 641-9 (1991).
387. Weinstein, P.P. & Jaffe, J.J. Cobalamin and folate metabolism in helminths. *Blood Rev* **1**, 245-53 (1987).
388. Rossi, M., Amaretti, A. & Raimondi, S. Folate production by probiotic bacteria. *Nutrients* **3**, 118-34 (2011).
389. 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 Negl Trop Dis* **6**, e1621 (2012).
390. James, C.E., Hudson, A.L. & Davey, M.W. Drug resistance mechanisms in helminths: is it survival of the fittest? *Trends Parasitol* **25**, 328-35 (2009).
391. Vercruysse, J., Levecke, B. & Prichard, R. Human soil-transmitted helminths: implications of mass drug administration. *Curr Opin Infect Dis* **25**, 703-8 (2012).
392. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* **42**, D1091-7 (2014).
393. Peixoto, C.A. & Silva, B.S. Anti-inflammatory effects of diethylcarbamazine: a review. *Eur J Pharmacol* **734**, 35-41 (2014).
394. Martin, R.J., Robertson, A.P. & Bjorn, H. Target sites of anthelmintics. *Parasitology* **114 Suppl**, S111-24 (1997).
395. Jex, A.R. *et al.* *Ascaris suum* draft genome. *Nature* **479**, 529-33 (2011).
396. Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S. & Hopkins, A.L. Quantifying the chemical beauty of drugs. *Nat Chem* **4**, 90-8 (2012).
397. Pensel, P.E., Albani, C., Gamboa, G.U., Benoit, J.P. & Elisondo, M.C. In vitro effect of 5-fluorouracil and paclitaxel on *Echinococcus granulosus* larvae and cells. *Acta Trop* **140**, 1-9 (2014).

Supplementary Figures



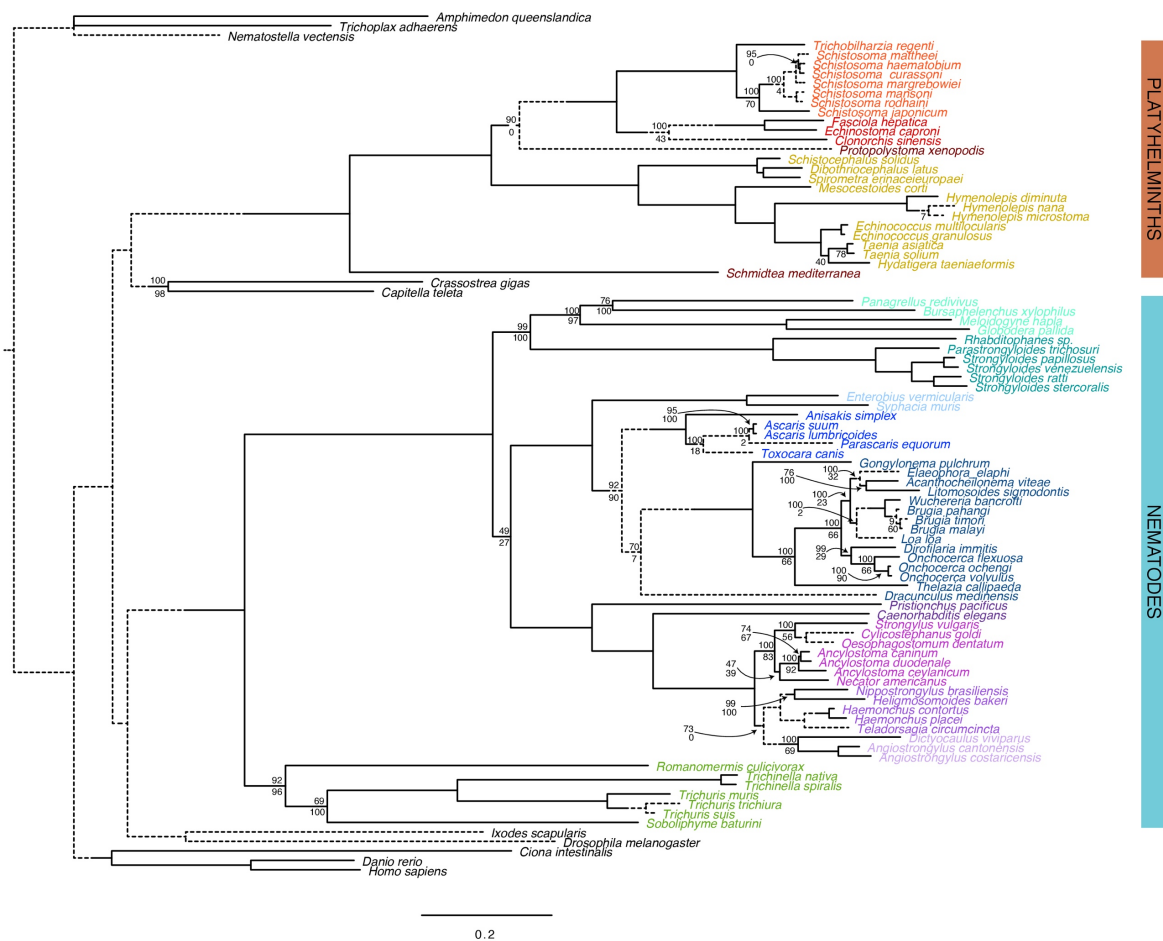
Supplementary Figure 1. Assembly and gene set statistics.

Blue rows indicate the 33 'tier 1' genomes. 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).



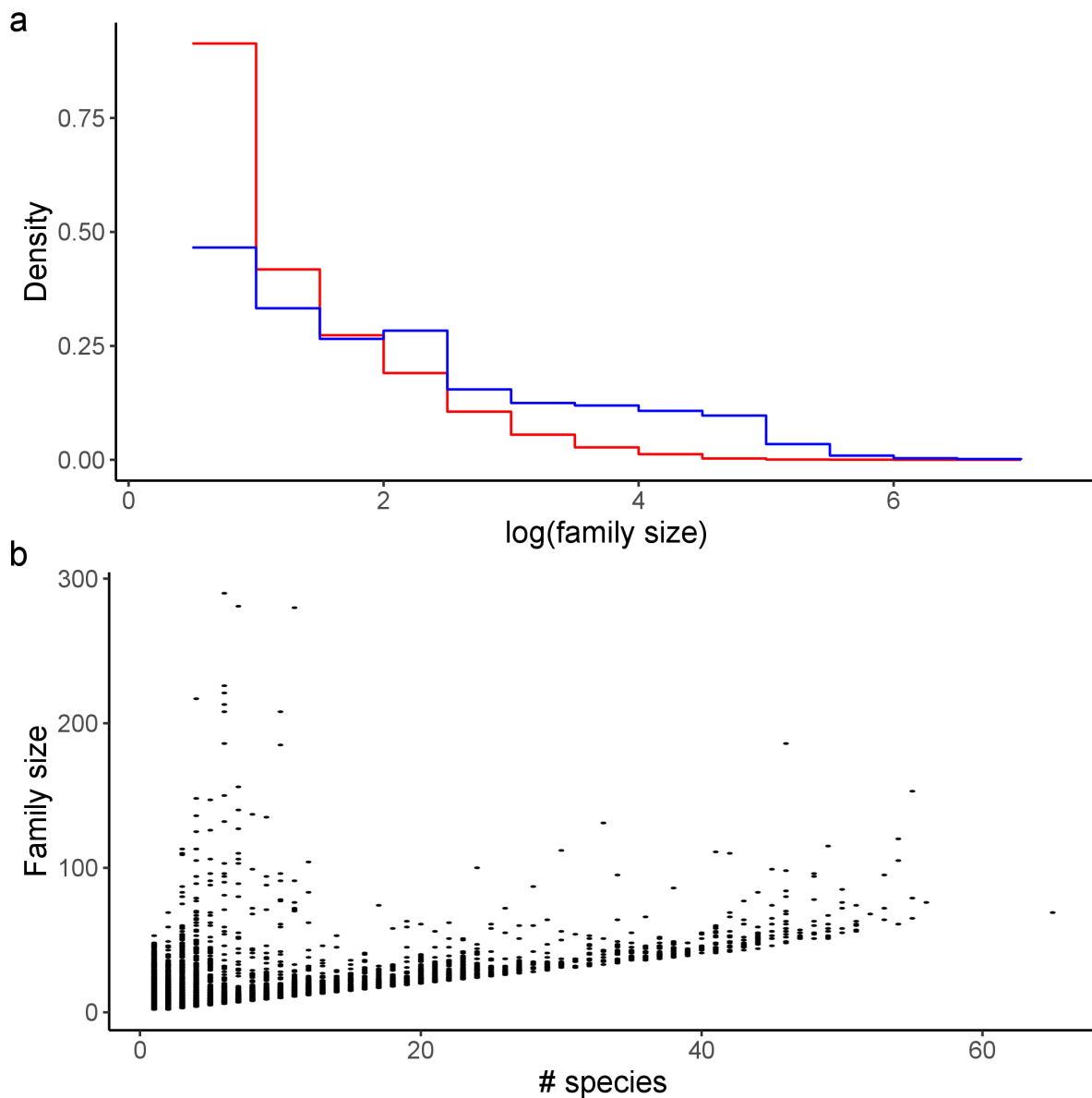
Supplementary Figure 2. Patterns of gene family sharing.

(a) The numbers of gene families (pink bars; values on left-hand y-axis) and the numbers of genes in those families (blue bars; values on right-hand y-axis) with particular patterns of sharing between high-level groups in our Compara data. Shading in the lower panel from pink to blue represents how widespread each set of families are, with pink representing families specific to one group and dark blue those families present in all groups. (b) Scatterplot of gene family size against the number of species a family is present in, with each point representing a single gene family (families with less than 3 genes are excluded), and points coloured according to the number of higher-level taxonomic groups they are shared between, as in the lower part of panel (a).



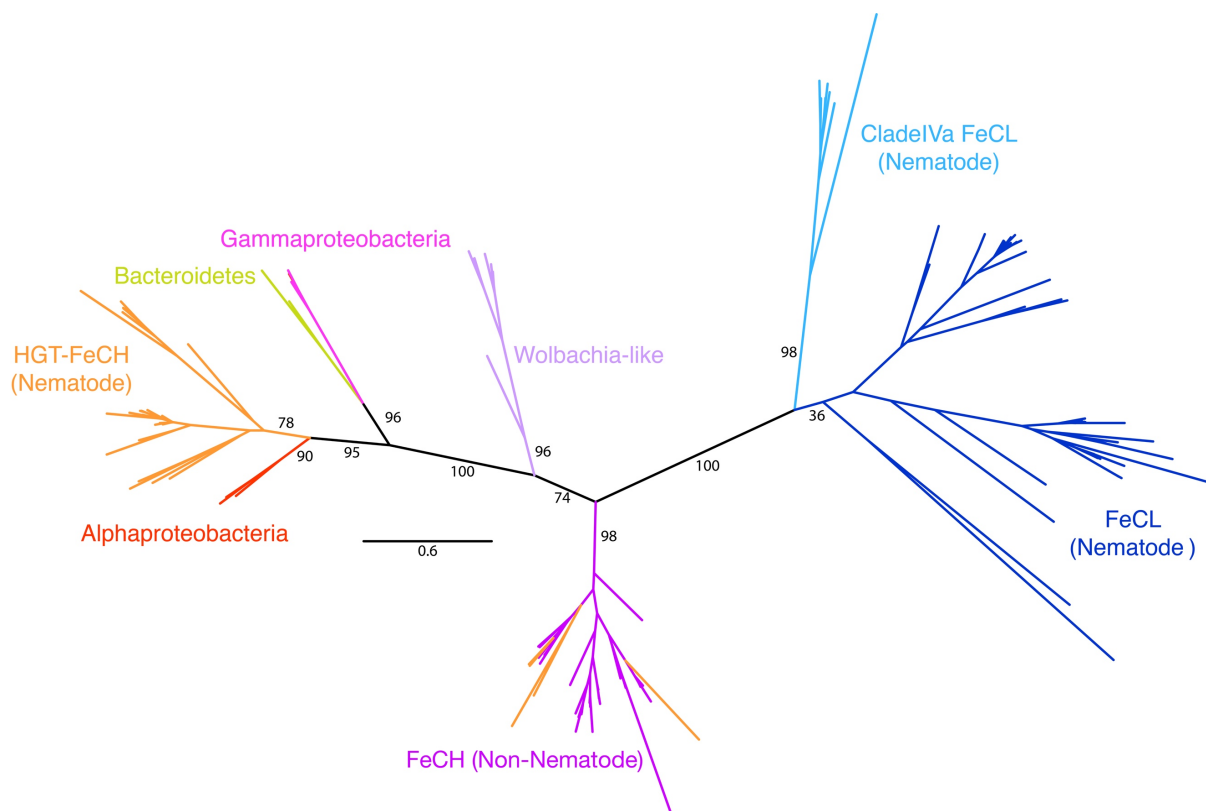
Supplementary Figure 3. Support for species tree relationships by gene family content.

The phylogenetic tree shown is the species tree from single-copy families, also shown in Fig. 1 (Methods 5). 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 Spiruromorpha+other, Ascaridomorpha and Oxyuridomorpha within clade III; and the positions of the genera *Bursaphelenchus* and *Dictyocaulus* within Nematoda (Supplementary Results 2.2).



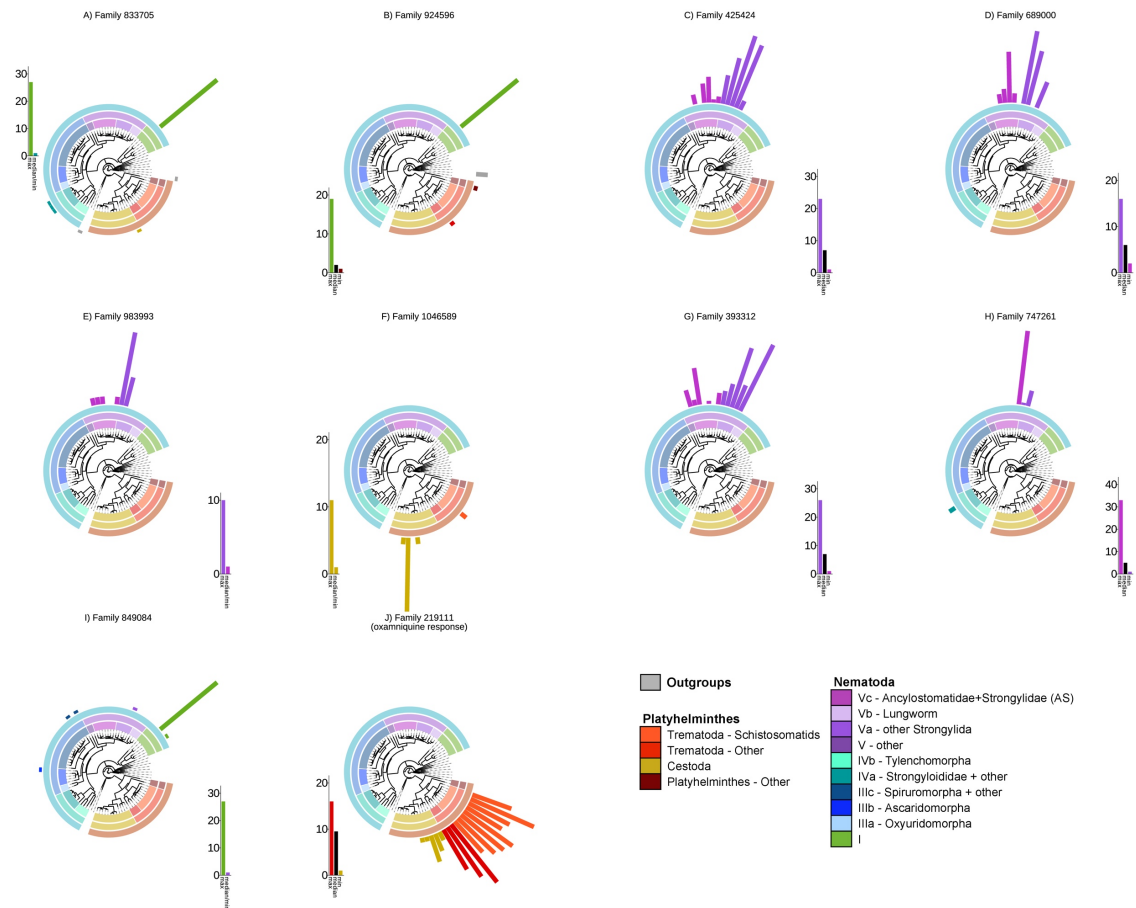
Supplementary Figure 4. Hypothetical proteins.

(a) Histograms of log(family size) for the Compara gene families that lacked functional annotation (protein names; Supplementary Results 2.1) and families with annotations (red and blue lines, respectively). Half of the Compara gene families (46.9%) lacked functional annotation. These families tended to be smaller than those with annotations ($P < 2.2 \times 10^{-16}$, Kolmogorov-Smirnov test). (b) Size distribution of hypothetical families shows that many were found in a large number of species, suggesting that they contained genuine genes of unknown function.



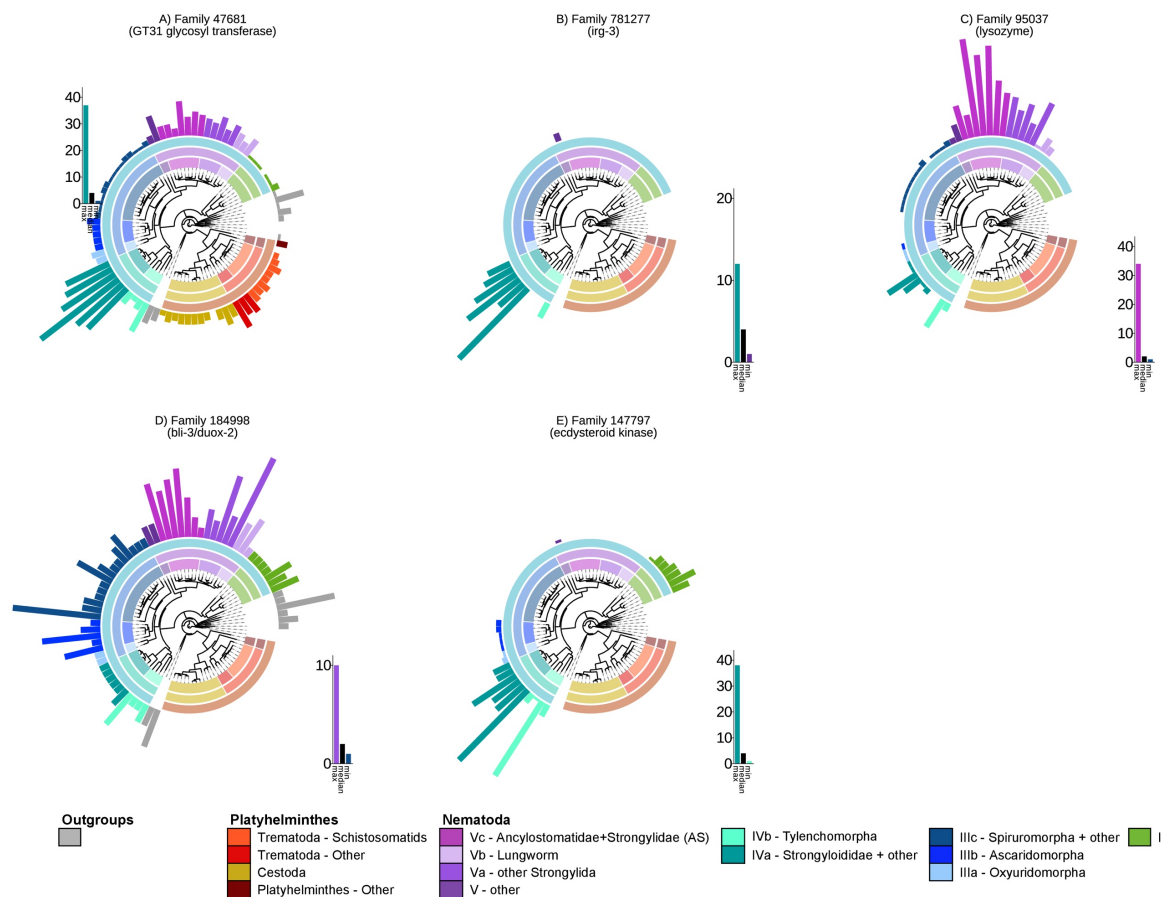
Supplementary Figure 5. Phylogenetic tree of ferrochelataes.

Branches are coloured by membership of protein sequences to ferrochelatae Compara families. Bootstrap support of the main branches is indicated. FeCL (nematode): nematode specific non-functional ferrochelatae-like proteins, devoid of active site (family 740872); CladeIVa FeCL (nematode): synapomorphic non-functional ferrochelatae-like family of nematode clade IVa (family 1184543); FeCH (Non-nematode): functional ferrochelatae family composed of taxa of non-nematode phyla (family 850580 and part of family 787620); Alphaproteobacteria: ferrochelataes of Alphaproteobacteria (Rhizobiales); Bacteroidetes: ferrochelataes of *Leadbetterella byssophila* and *Mucilaginibacter paludis*; Gammaproteobacteria: ferrochelataes of *Pseudomonas* species; Wolbachia-like: ferrochelataes of *Wolbachia*, *Ehrlichia* and *Hydrogenobacter* species; and HGT-FeCH (nematode): clade III/IV specific functional ferrochelatae acquired through horizontal gene transfer from Alphaproteobacteria (family 787620).



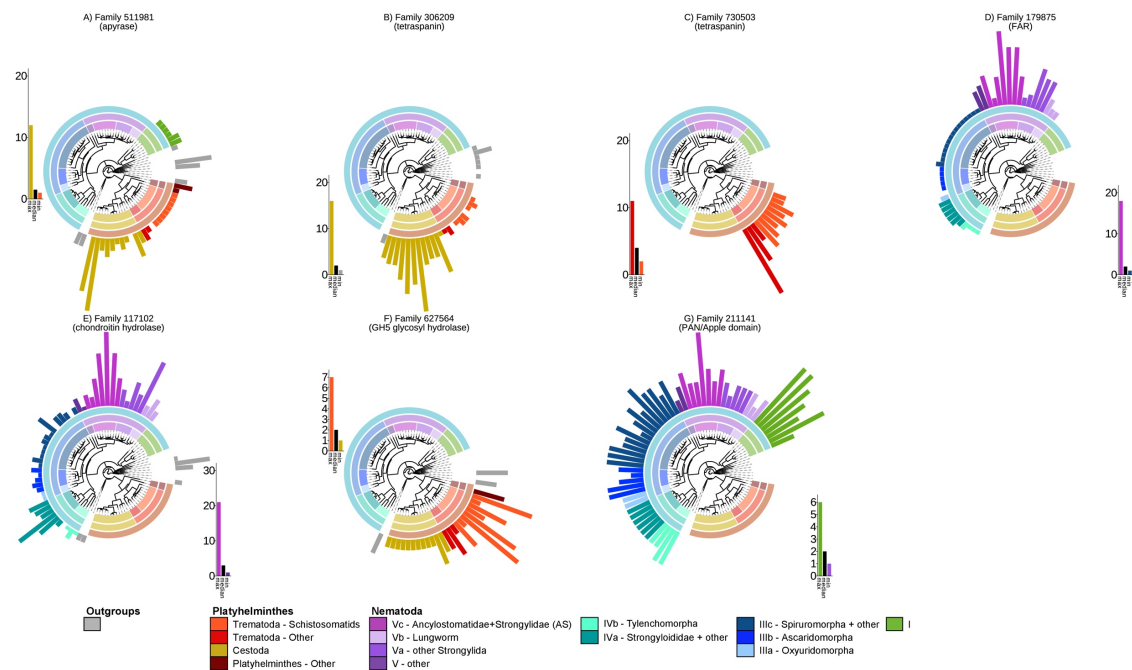
Supplementary Figure 7. Striking expansions in families with poorly defined roles.

Gene families with striking variation across species but uncharacterised (a-i), or poorly characterised (j) functions. 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.



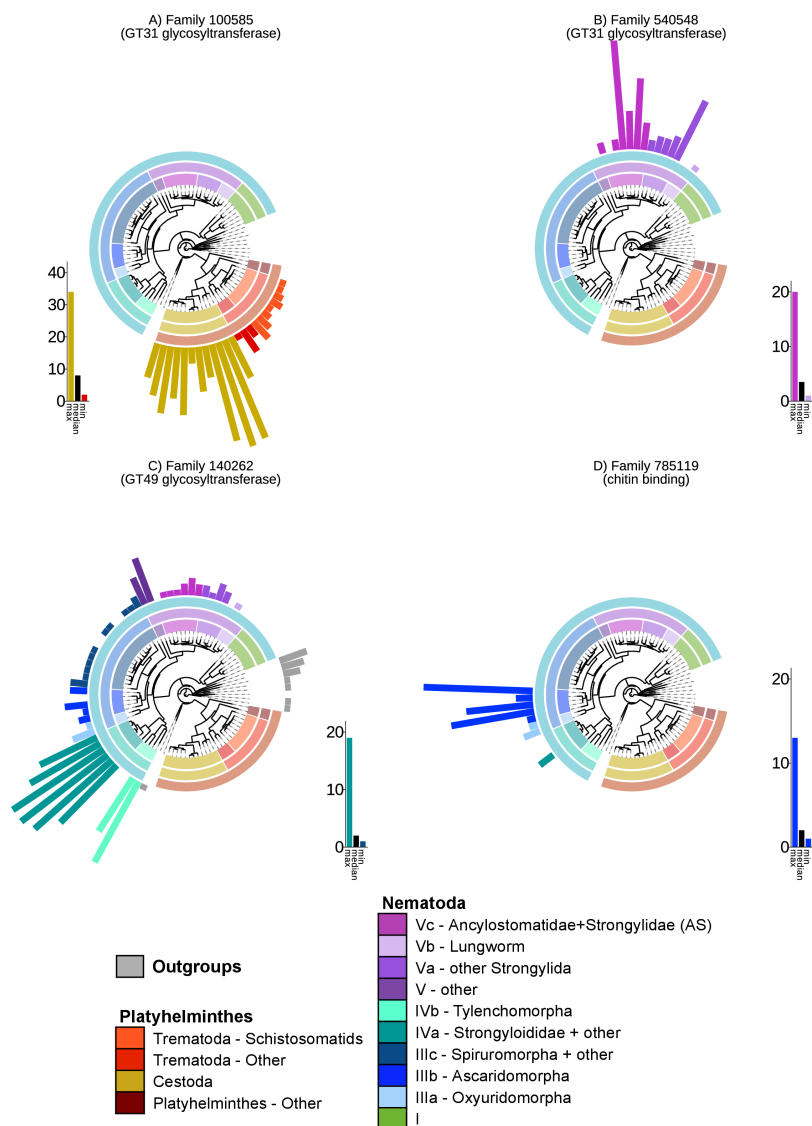
Supplementary Figure 8. Expanded families involved in immunity and development.

Expanded gene families with potential roles in immunity and development. 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.



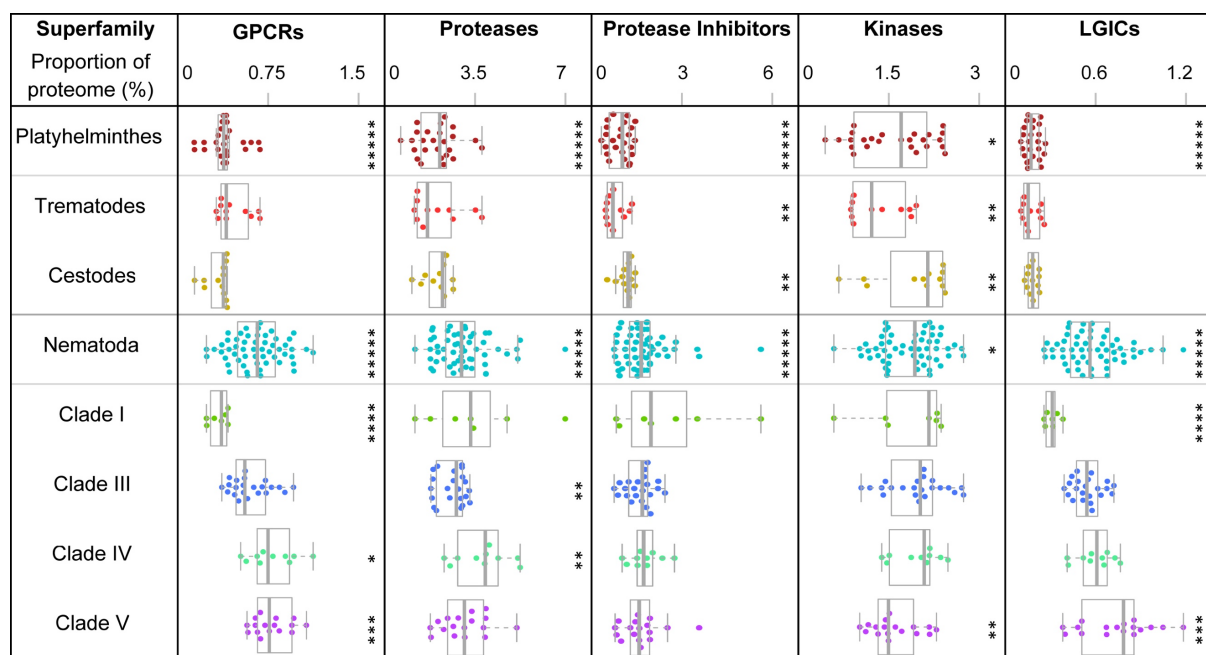
Supplementary Figure 9. Expanded families involved in parasitism.

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.



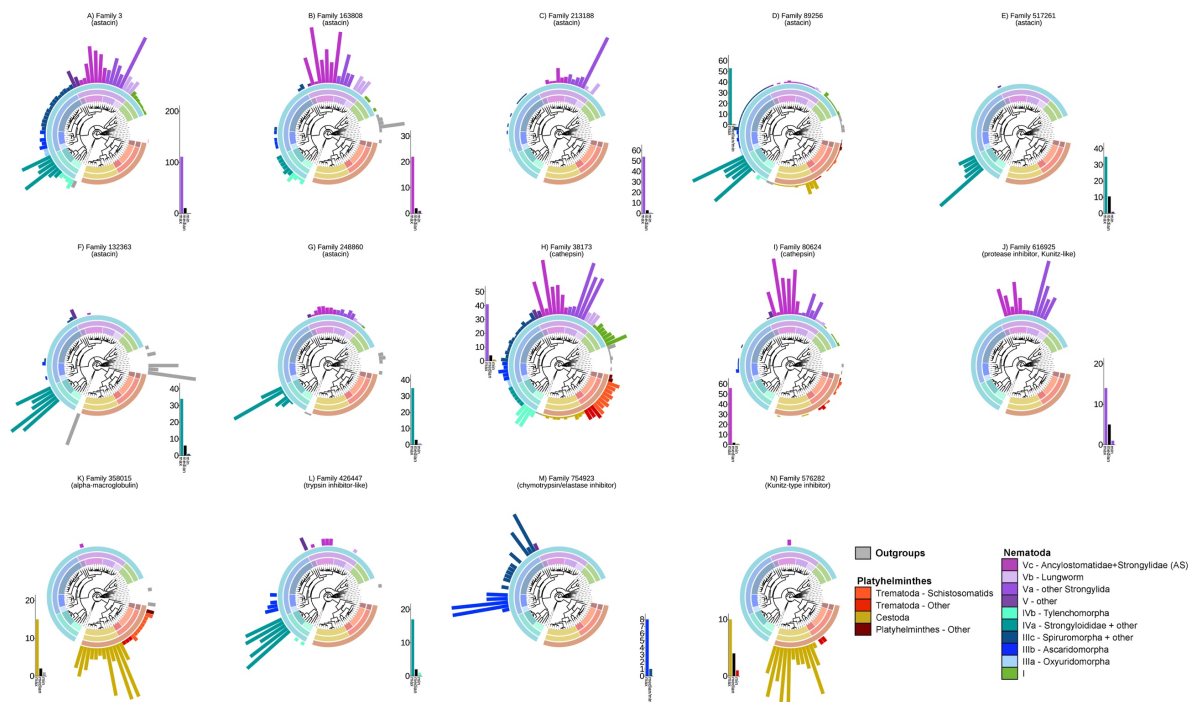
Supplementary Figure 10. Surface coat expanded families.

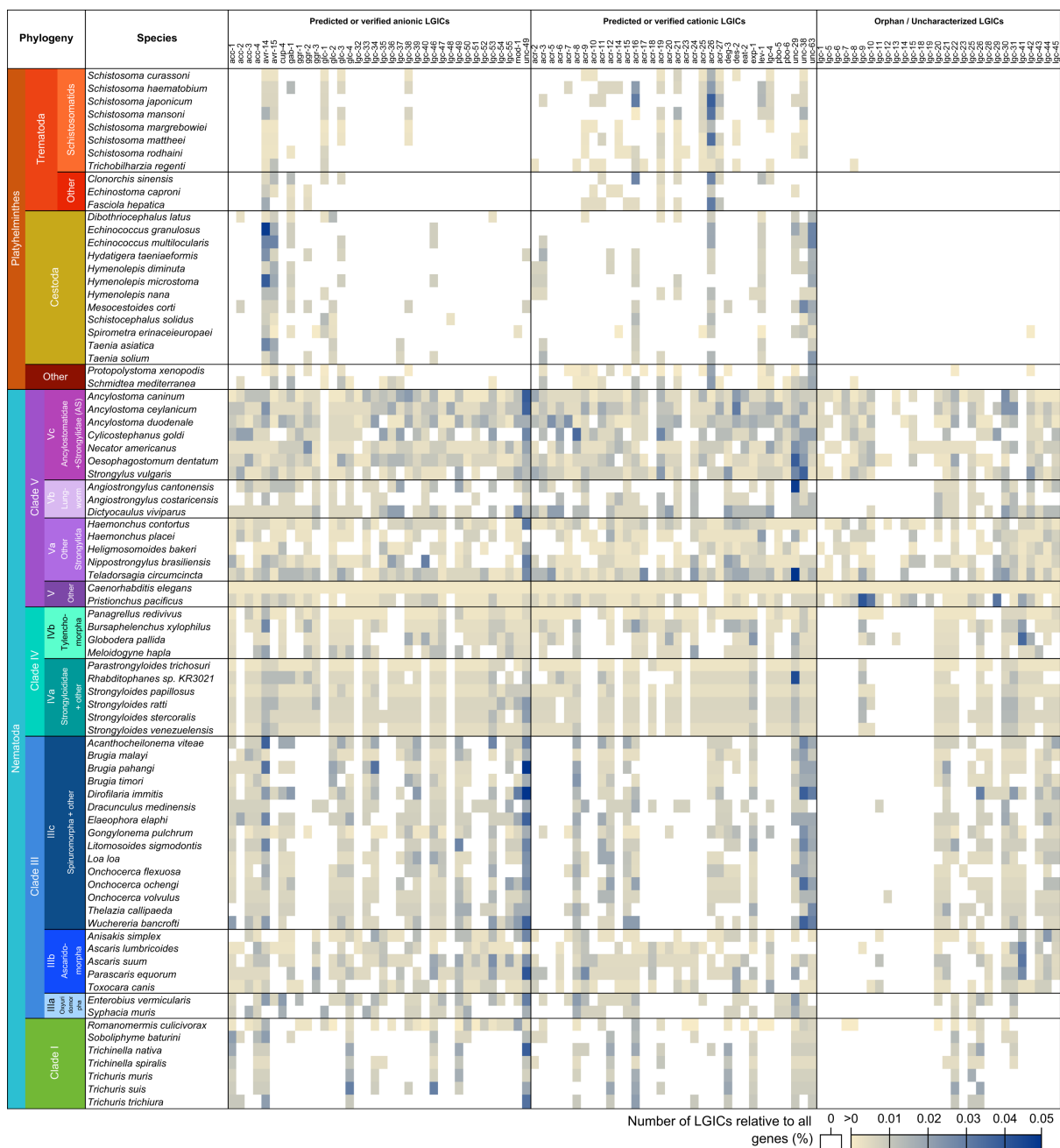
Expanded families that may be involved in the surface coat. The plot for a family shows the gene count in each species, superimposed upon our species tree. A scale bar beside the plot for each family shows the minimum, median, and maximum gene count across the species, for that family.



Supplementary Figure 13. Abundances of superfamilies historically targeted for drug development.

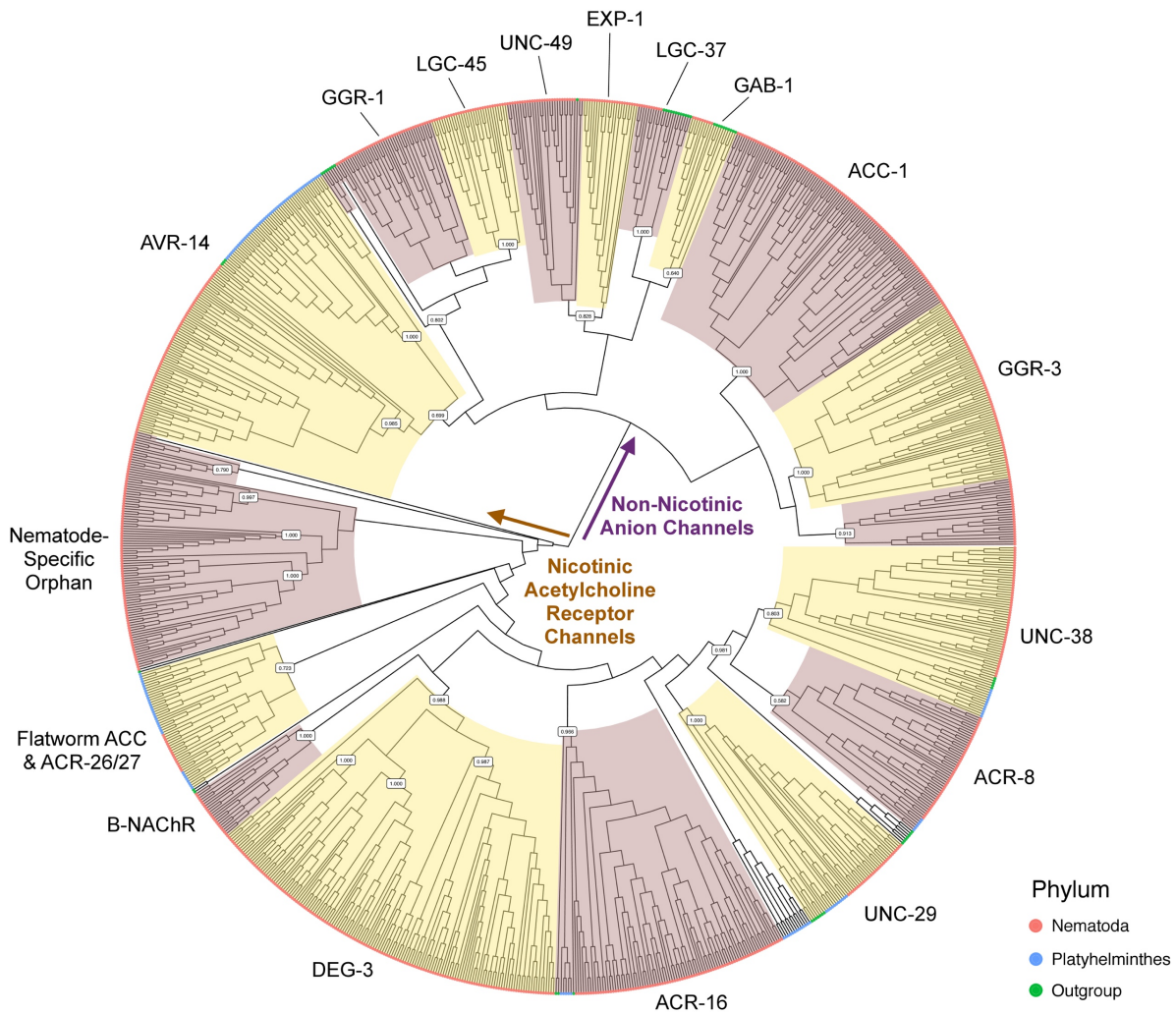
(a) Abundances (relative to the total numbers of predicted genes) of genes belonging to superfamilies of interest. Comparisons are (i) platyhelminths vs nematodes, (ii) trematodes vs cestodes and (iii) nematode clade vs other nematodes. * $P < 0.05$, ** $P < 0.01$, *** $P < 10^{-3}$, **** $P < 10^{-4}$, ***** $P < 10^{-5}$ (Mann Whitney U test). The *C. elegans* GPCR value of 4.6% was not shown. The left and right of a box are the lower and upper quartiles, respectively, while the line inside the box is the median. The whiskers represent the most extreme data points in the range $Q1 - (1.5 \times IQR)$ to $Q3 + (1.5 \times IQR)$. The individual data points are displayed as a swarmplot.





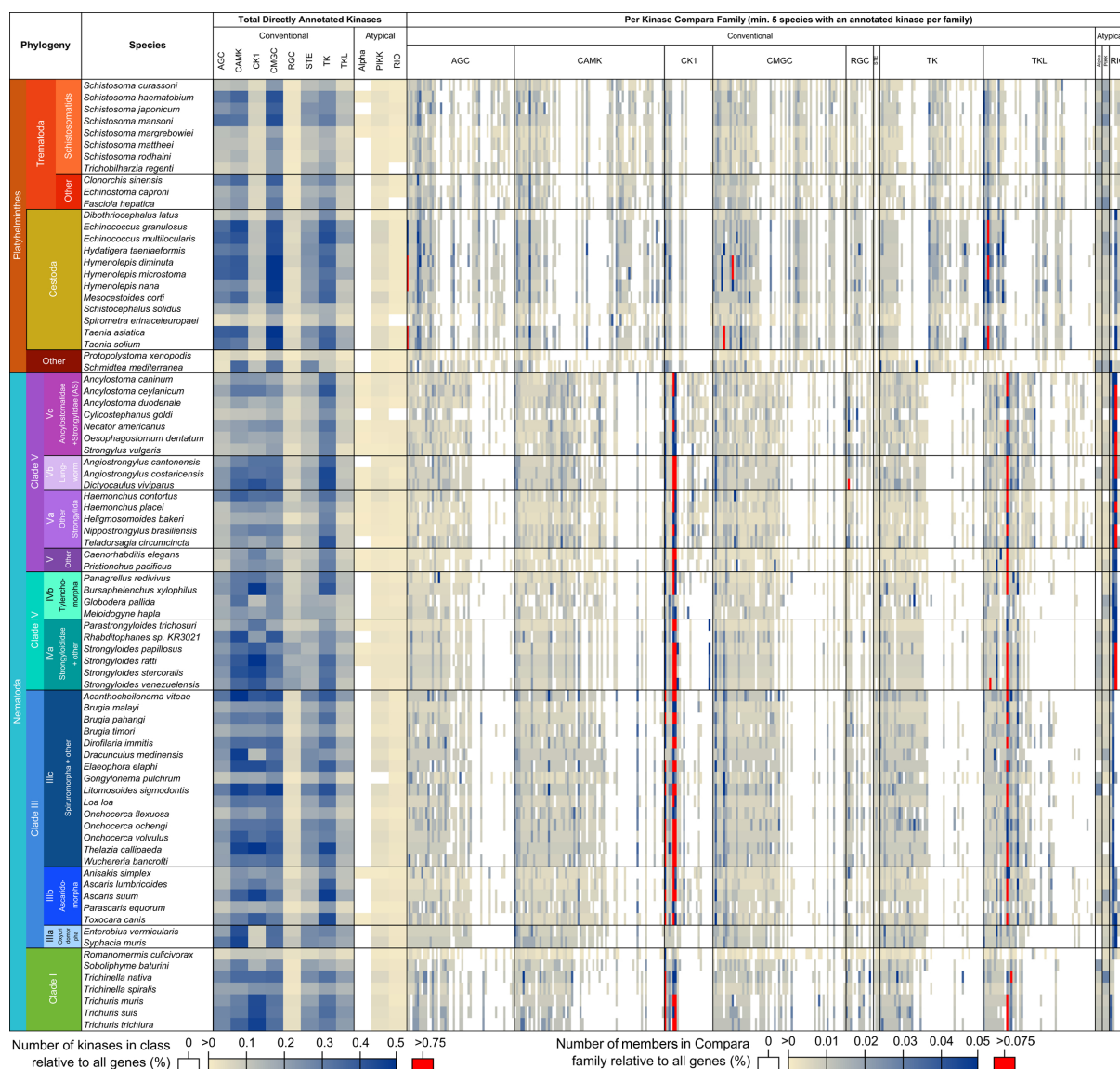
Supplementary Figure 16. Heatmap of ligand gated ion channels.

Relative abundance profiles for all 99 LGICs represented in at least 3 of the 81 helminth species (Methods 11). 3 LGICs present in fewer than 3 species were omitted from the visualisation. For each species, the gene count in a class was normalised by dividing by the total gene count for that species.



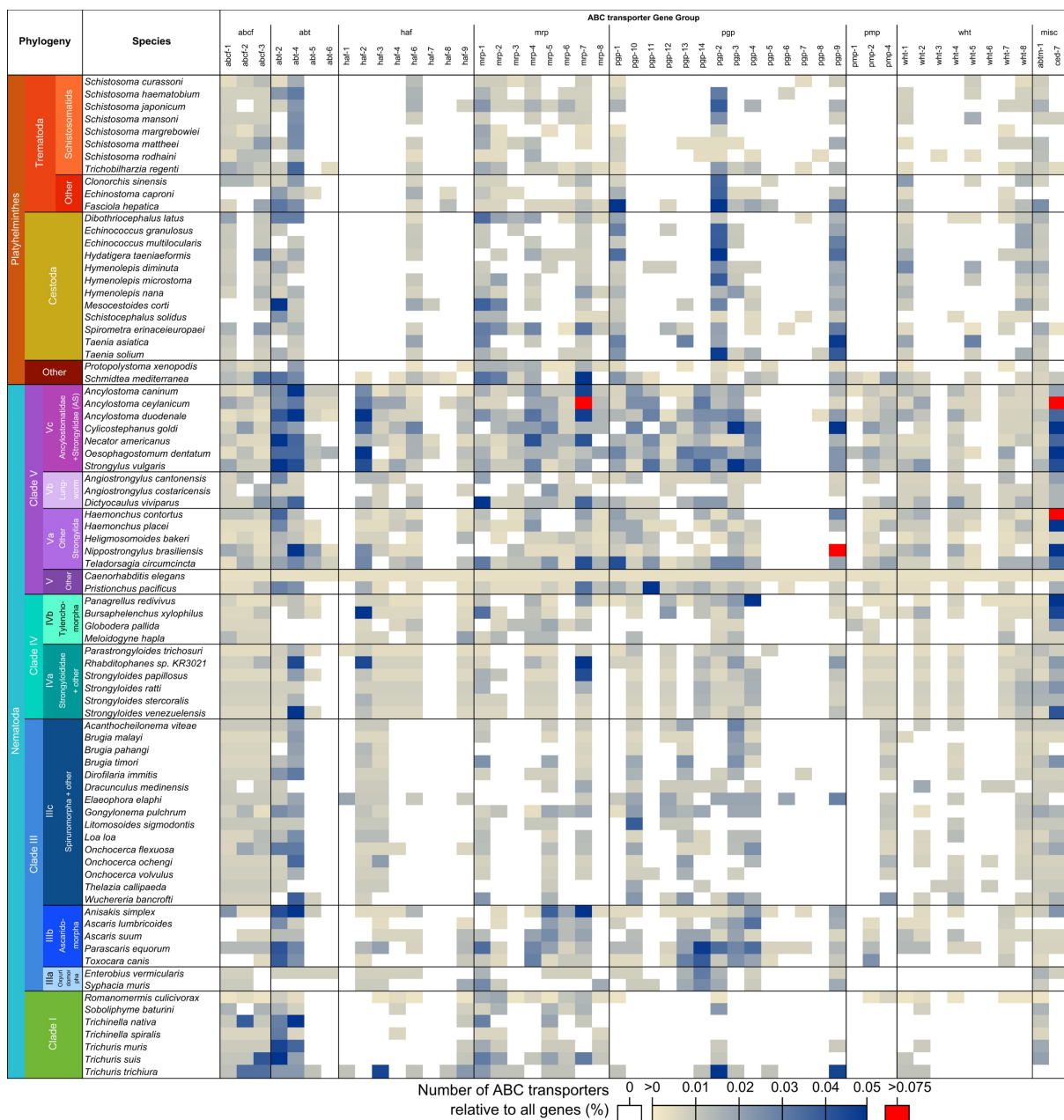
Supplementary Figure 17. Inferred phylogeny of the cys-loop superfamily in platyhelminths and nematodes.

Posterior probabilities were calculated from 8 reversible jump MCMC chains using MrBayes. The tree is rooted between nicotinic acetylcholine receptors and non-nicotinic anion channels. Posterior probabilities are displayed for nodes that correspond to our classification of cys-loop proteins.



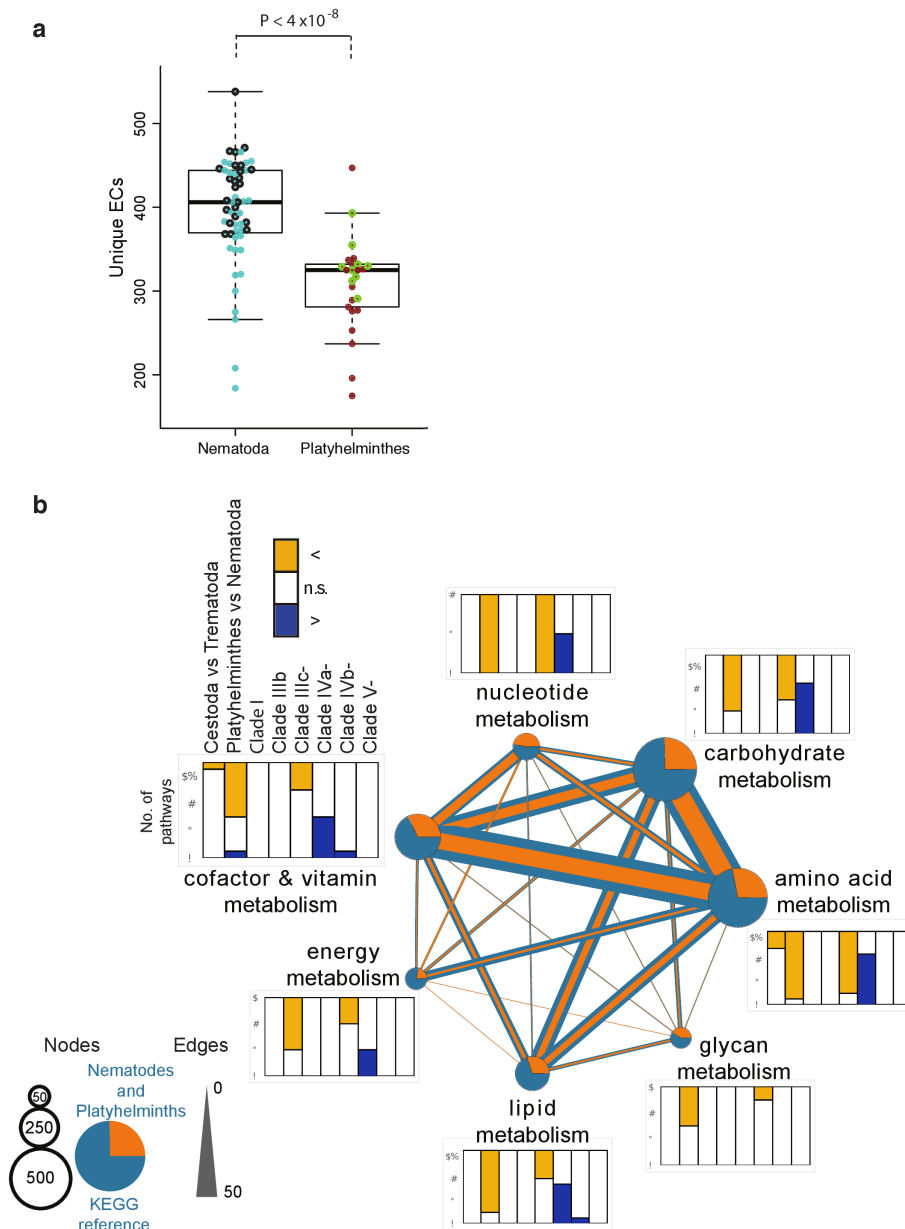
Supplementary Figure 18. Kinases heatmap.

Relative abundance profiles for all kinase genes, and for 343 individual kinase Compara gene families represented in at least 5 of the 81 helminth species (Methods 11). 689 Compara families present in fewer than 5 species and an additional 43 unclassified kinase families were omitted from the visualisation. For each species, the gene count in a class was normalised by dividing by the total gene count for that species. 'Total Directly Annotated Kinases' represents the relative abundance of kinases that were directly annotated per species, while 'Per Kinase Compara Family' represents the relative abundance of genes per annotated kinase family.



Supplementary Figure 19. Heatmap of ABC transporters.

Relative abundance profiles for 50 ABC transporter classes (Methods 11). For each species, the gene count in a class was normalised by dividing by the total gene count for that species.



Supplementary Figure 20. Comparison of EC number annotations and KEGG pathway coverage.

(a) Unique enzymes (EC numbers) annotated before using pathway hole-filling. The box represents the range between the first and third quartiles, with the median shown as a bold line inside the box. The whiskers represent the most extreme data points in the range $Q1-(1.5 \times IQR)$ to $Q3+(1.5 \times IQR)$. Tier 1 species are highlighted with thick borders (black border for Nematoda and green border for Platyhelminthes). (b) Comparison of KEGG reference pathway coverages (Methods 19), 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).

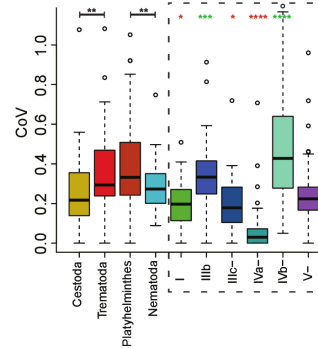
a

C	Cestoda
T	Trematoda
P	Platyhelminthes
N	Nematoda
I	Clade I Nematoda
IIIb	Clade IIIb Nematoda
IIIc-	Platylael Nematoda (Clade IIIc, except 3 non filarial worms)
IVa-	Clade IVa Nematoda (except free living <i>Rhabdiostrongylus</i> sp. J02021)
IVb-	Clade IVb Nematoda (except free living <i>P. redivivus</i>)
V-	Clade V Nematoda (except free living <i>C. elegans</i> and <i>P. pacificus</i>)

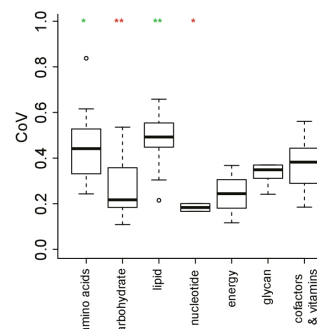


Superpathways	Pathways	C vs T	P vs N	I	IIIb	IIIc-	IVa-	IVb-	V-
Amino acid metabolism	Alanine, aspartate and glutamate metabolism								
	Arginine and proline metabolism								
	Cysteine and methionine metabolism								
	Glycine, serine and threonine metabolism								
	Histidine metabolism								
	Lysine biosynthesis								
	Lysine degradation								
	Phenylalanine metabolism								
	Phenylalanine, tyrosine and tryptophan biosynthesis								
	Tryptophan metabolism								
	Tyrosine metabolism								
	Valine, leucine and isoleucine biosynthesis								
Carbohydrate metabolism	Valine, leucine and isoleucine degradation								
	Amino sugar and nucleotide sugar metabolism								
	Ascorbate and aldarate metabolism								
	Butanoate metabolism								
	Citrate cycle								
	Fructose and mannose metabolism								
	Galactose metabolism								
	Glycolysis / Gluconeogenesis								
	Glyoxylate and dicarboxylate metabolism								
	Inositol phosphate metabolism								
	Pentose and glucuronate interconversions								
	Pentose phosphate pathway								
Lipid metabolism	Propanoate metabolism								
	Pyruvate metabolism								
	Starch and sucrose metabolism								
	alpha-Linolenic acid metabolism								
	Arachidonic acid metabolism								
	Biosynthesis of unsaturated fatty acids								
	Ether lipid metabolism								
	Fatty acid biosynthesis								
	Fatty acid elongation in mitochondria								
	Fatty acid metabolism								
	Glycerolipid metabolism								
	Glycerophospholipid metabolism								
Nucleotide metabolism	Linoleic acid metabolism								
	Sphingolipid metabolism								
	Steroid biosynthesis								
	Synthesis and degradation of ketone bodies								
	Purine metabolism								
	Pyrimidine metabolism								
	Nitrogen metabolism								
	Oxidative phosphorylation								
	Sulfur metabolism								
	Glycosaminoglycan degradation								
	Glycosphingolipid biosynthesis - ganglio series								
	Glycosphingolipid biosynthesis - globo series								
Metabolism of cofactors and vitamins	N-Glycan biosynthesis								
	O-Glycan biosynthesis								
	Other glycan degradation								
	Folate biosynthesis								
	Lipoic acid metabolism								
	One carbon pool by folate								
	Pantothenate and CoA biosynthesis								
	Porphyry and chlorophyll metabolism								
	Retinol metabolism								
	Riboflavin metabolism								
	Thiamine metabolism								
	Vitamin B6 metabolism								
	Nicotinate and nicotinamide metabolism								
	beta-Alanine metabolism								
	Cyanoamino acid metabolism								
	Glutathione metabolism								
	Selenoamino acid metabolism								

b



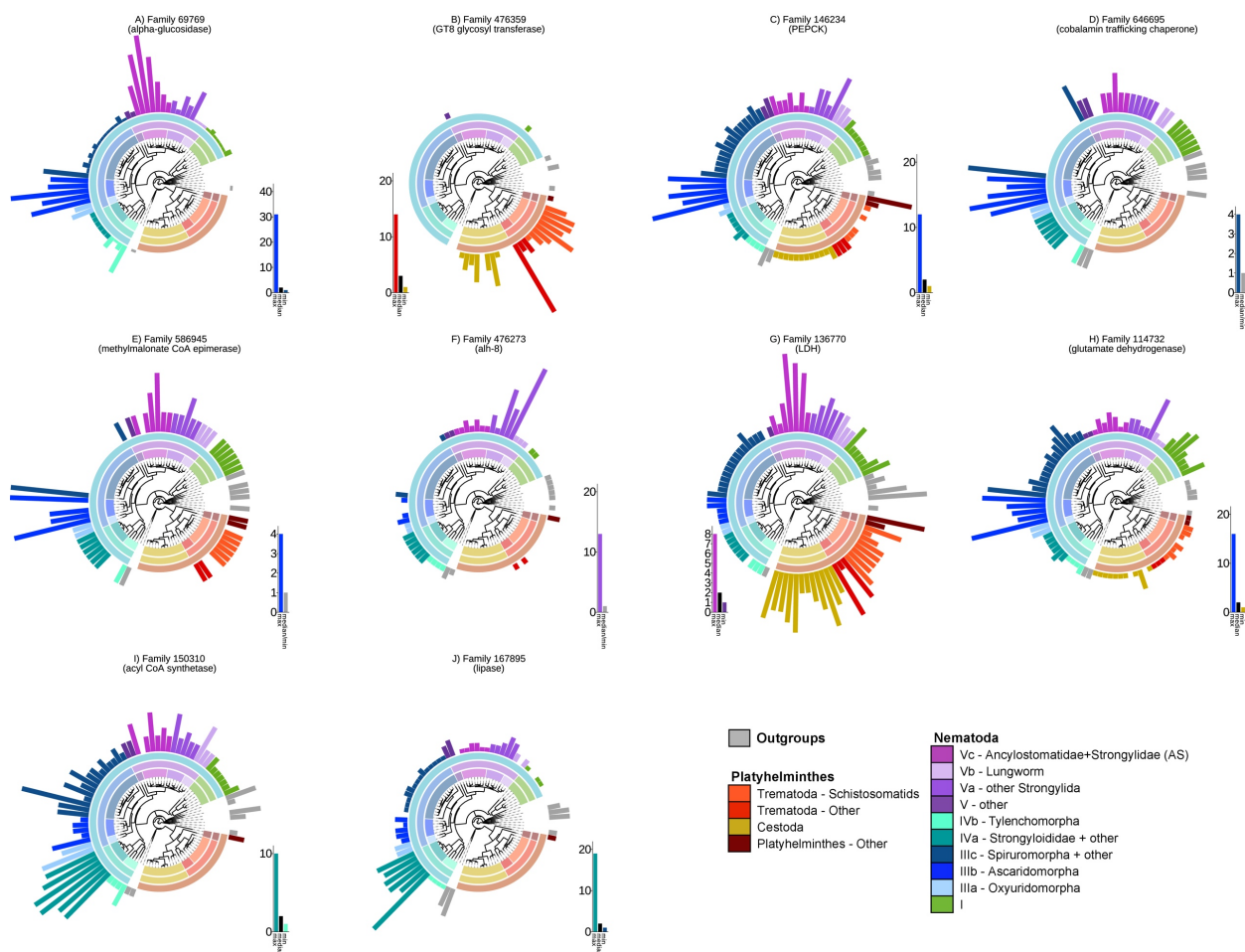
c



Supplementary Figure 21. Comparison of pathway coverage and variation.

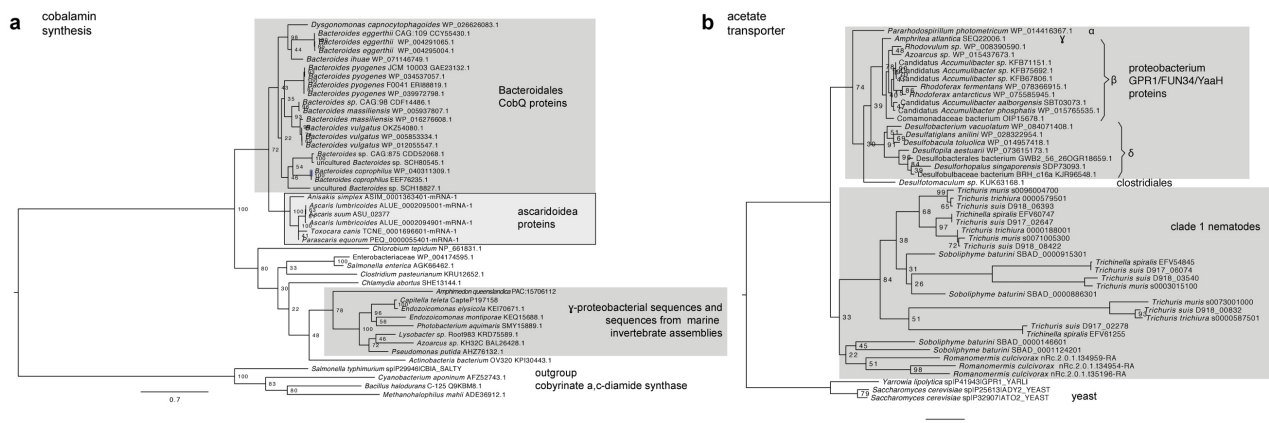
(a) Taxonomic differences in KEGG reference pathway coverage. For nematode clades, the comparison is with the union of the other five nematode groups. Platyhelminthes, Clade IIIc- and Clade IVa- show the most consistent overall pattern. Comparisons with Wilcoxon test P-value <0.05 (FDR corrected using the Benjamini-Hochberg procedure) are considered significant. (b) Within-group variation in KEGG pathway coverage, and differences between groups. Only helminth-relevant KEGG metabolic pathways were considered (Methods 13). The y-axis is the coefficient of variation of the KEGG pathway coverage. For nematode clades, the comparison is with the union of the other five nematode groups. (c) Variation among nematodes and platyhelminths in KEGG pathway coverage aggregated according to superpathways. In panels (b)

and (c), statistically significant comparisons are shown (red = lower; green = higher): * $0.01 < P \leq 0.05$; ** $1e-5 < P \leq 0.01$; *** $1e-10 < P \leq 1e-5$; **** $P < 1e-10$. For panels (b) and (c), the box represents the range between the first and third quartiles, with the median shown as a bold line inside the box. The whiskers represent $Q1 - (1.5 * IQR)$ and $Q3 + (1.5 * IQR)$.



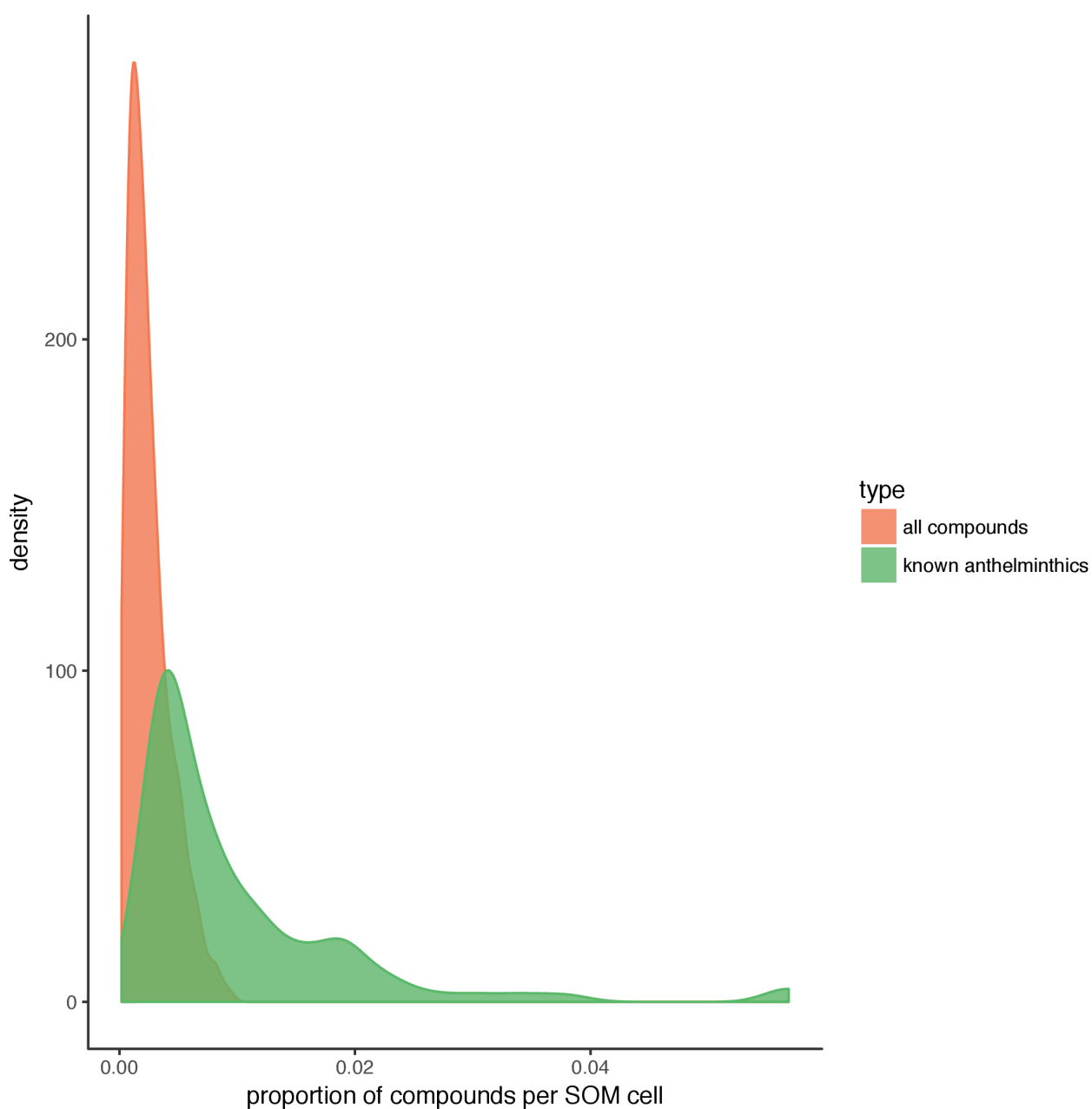
Supplementary Figure 22. Metabolism-related family expansions.

Expanded Compara families with metabolism-related functions. 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.



Supplementary Figure 23. Maximum likelihood phylogenies of two putative horizontal transfers.

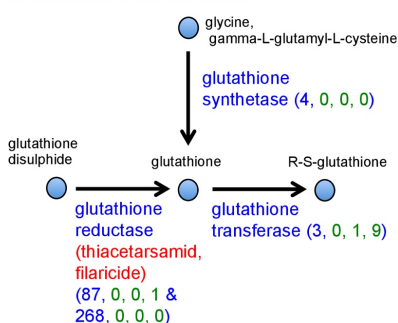
Phylogeny of helminth proteins, closely related sequences identified by BLAST and selected other sequences for (a) a clade IIIb-specific cobalamin-related family (CobQ/CbiP) gained from Bacteria, and (b) an acetate/succinate transporter in clade I nematodes that appears to have been gained from Bacteria and is likely to participate in acetate/succinate uptake or efflux.



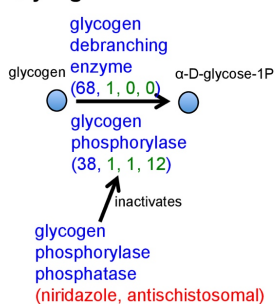
Supplementary Figure 24. Clustering of known anthelmintic compounds in our self-organising map.

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.

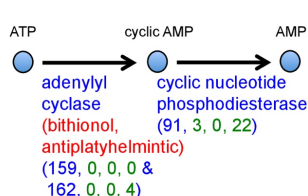
Glutathione metabolism



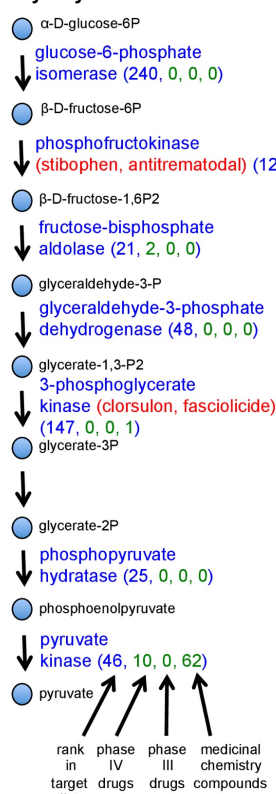
Glycogen metabolism



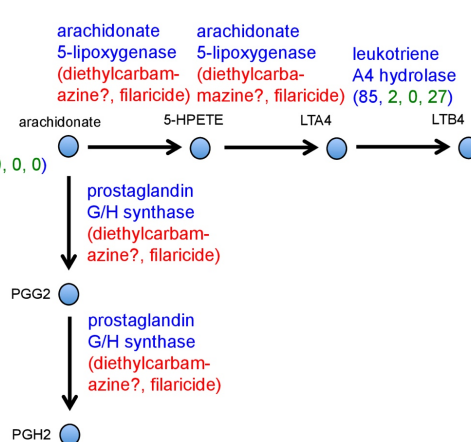
Purine metabolism



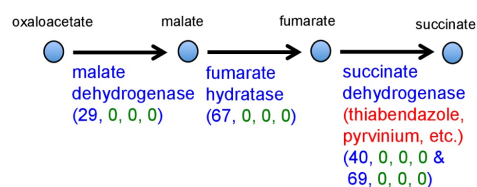
Glycolysis



Arachidonic acid metabolism

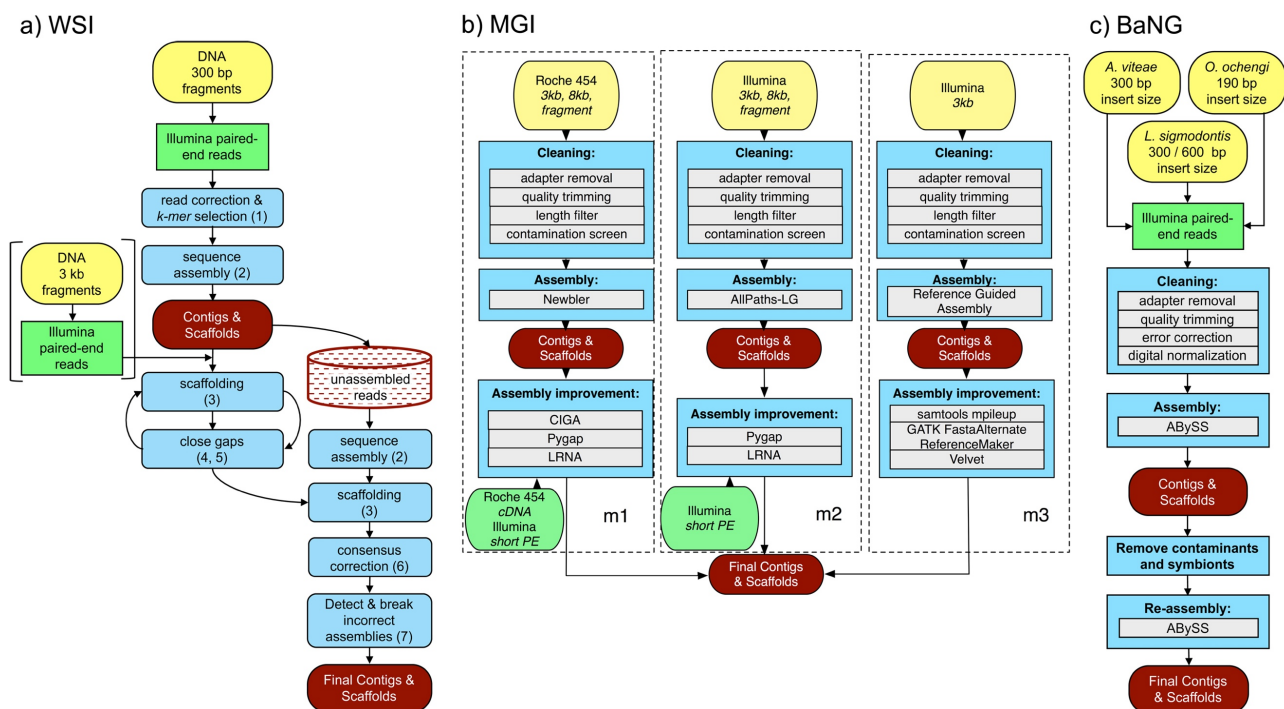


Malate dismutation



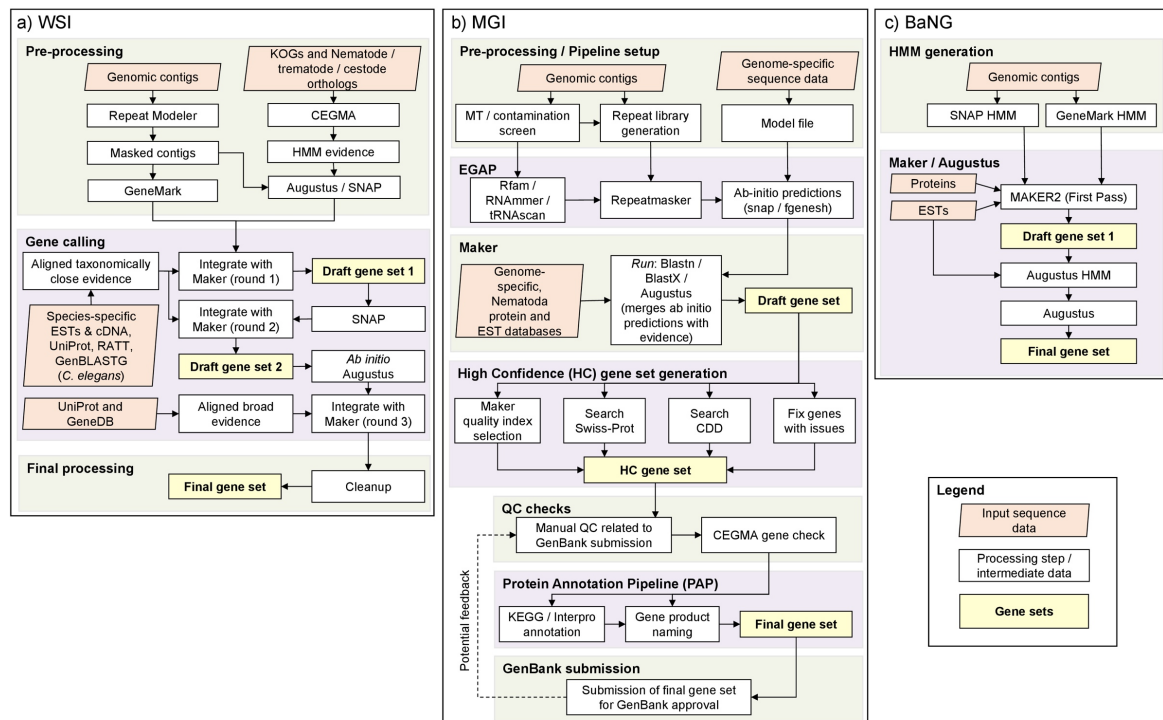
Supplementary Figure 25. Pathways with known targets.

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).



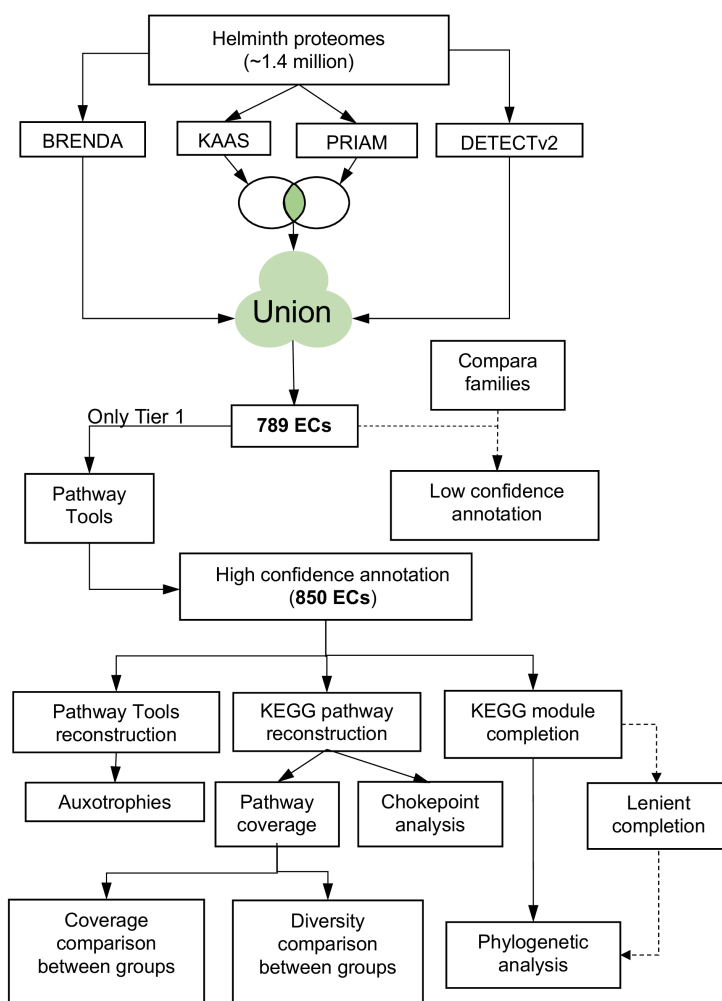
Supplementary Figure 26. Genome assembly pipelines.

The individual pipelines used (a) WSI (b) MGI and (c) BaNG. Details as defined in Methods.

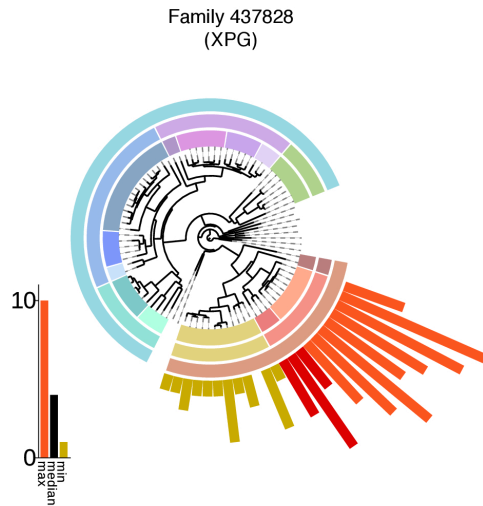


Supplementary Figure 27. Gene-finding pipelines.

The individual pipelines used at (a) WSI, (b) MGI and (c) BaNG. Details as defined in Methods.

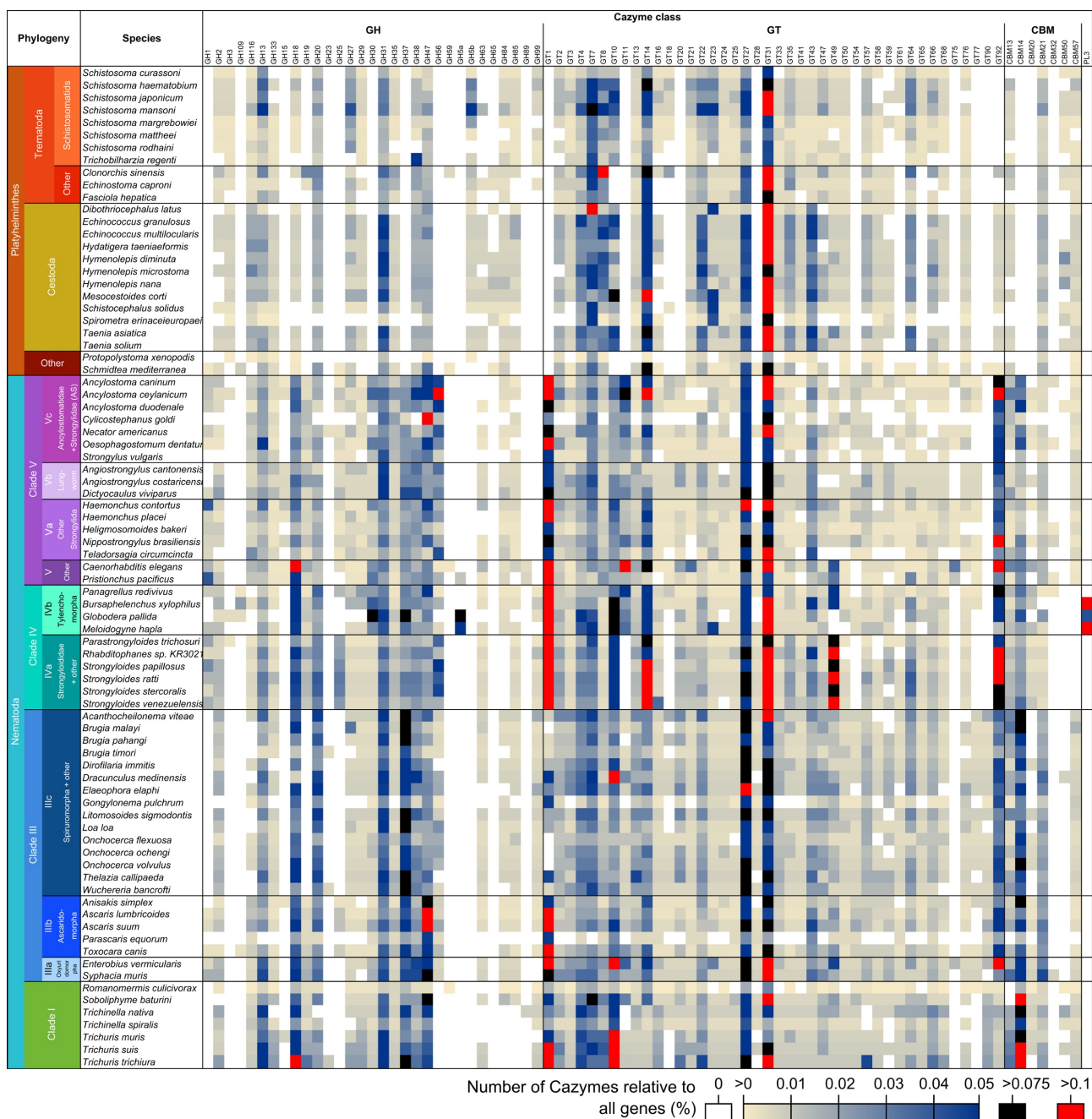


Supplementary Figure 28. Flowchart of methods for analysing metabolism.
Details as defined in Methods.



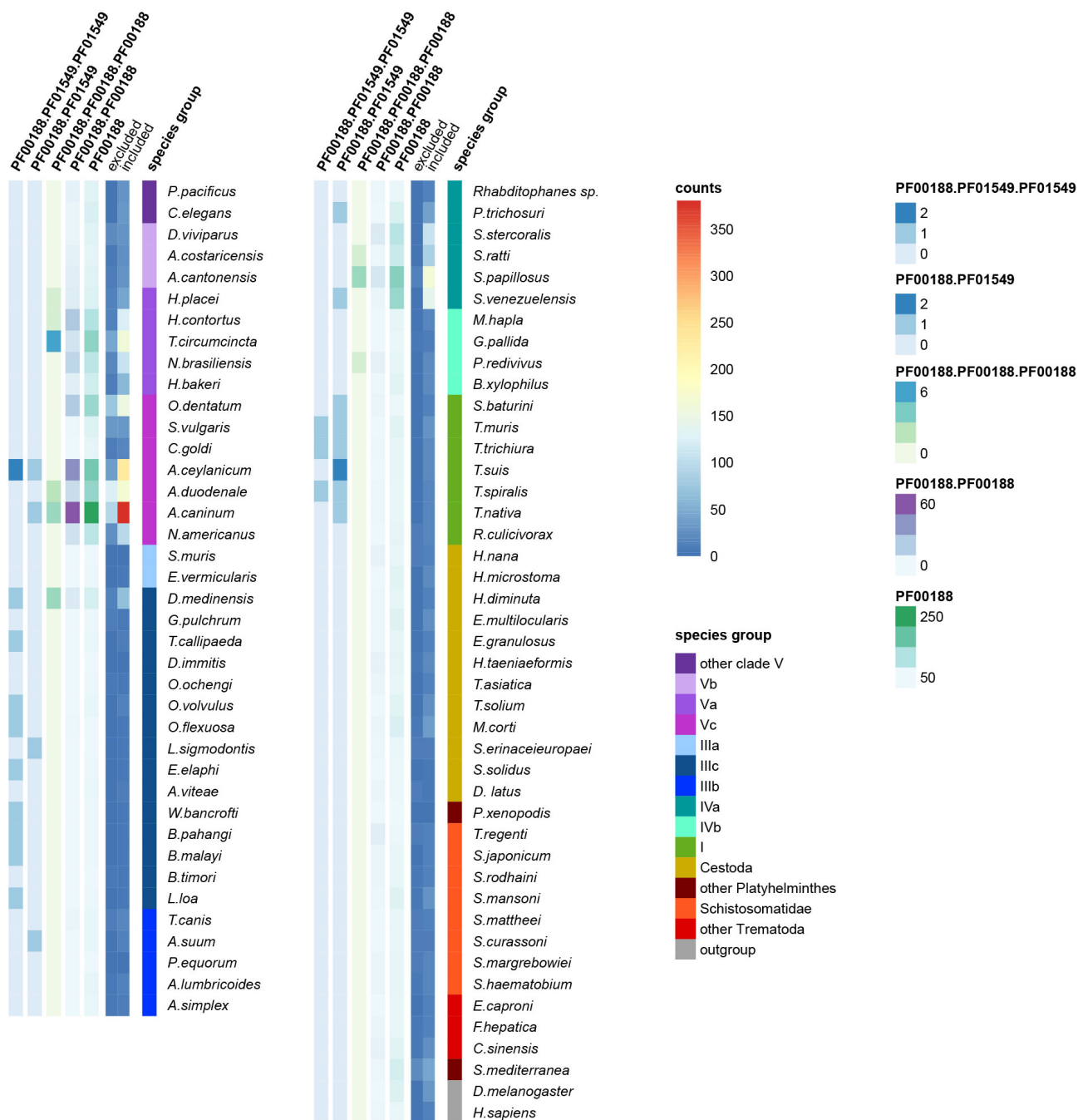
Supplementary Figure 29. Example of inexplicable expansion.

Expansion of the Compara XPG gene family. The plot shows the gene count in each species, superimposed upon our species tree. A scale bar beside the plot shows the minimum, median, and maximum gene count across the species, for that family.



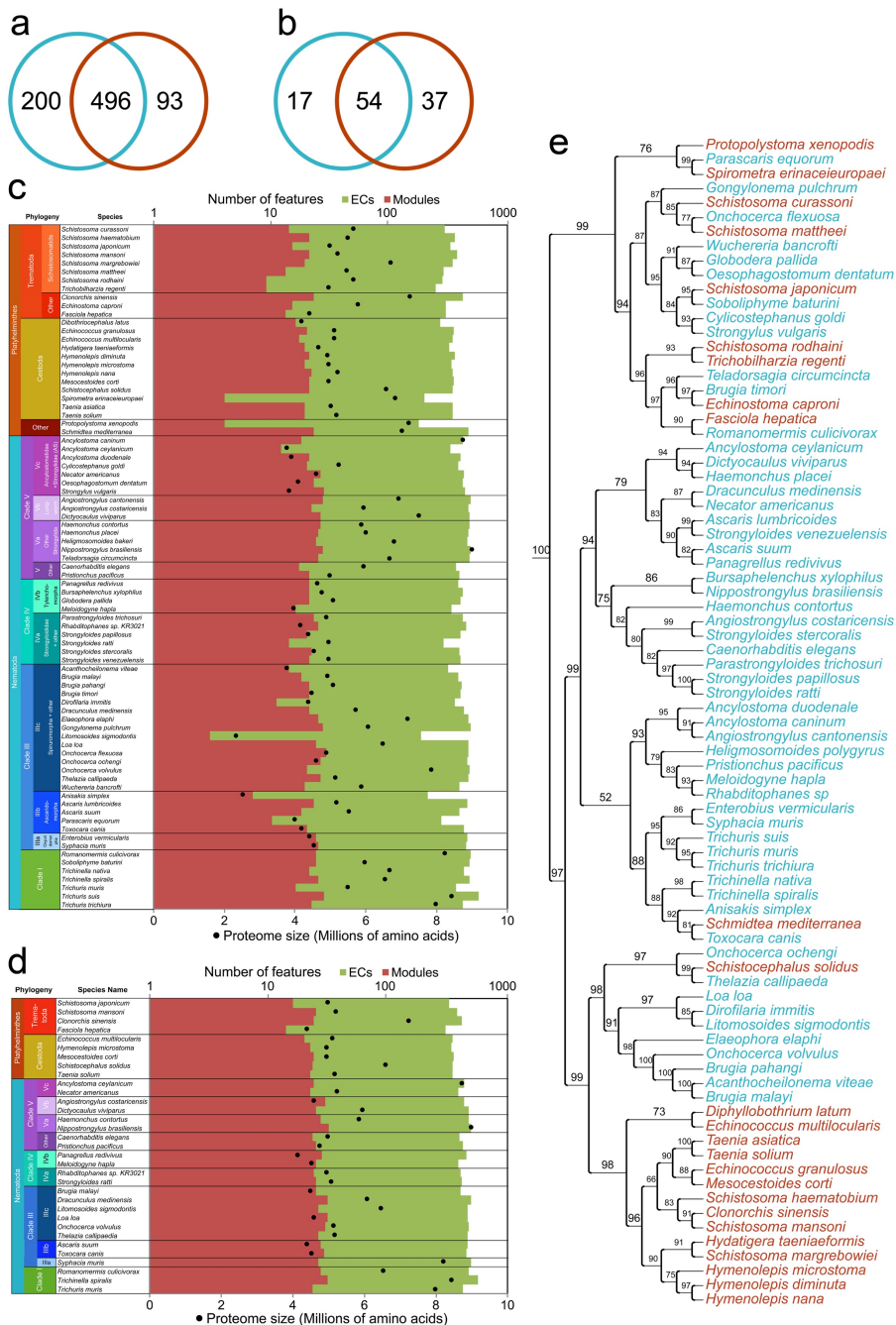
Supplementary Figure 30. CAZymes heatmap.

Relative abundance profiles for 81 CAZyme families represented in at least 3 of the 81 helminth species. 25 families present in fewer than 3 species were omitted from the visualisation. GH = Glycoside hydrolases, PL = Polysaccharide Lyases, GT = Glycosyltransferases, CBM = Carbohydrate-Binding Modules. For each species, the gene count in a class was normalised by dividing by the total gene count for that species.



Supplementary Figure 31. Domain composition of SCP/TAPS genes.

A domain combination is included in this matrix if the combination can be detected in more than five sequences amongst all species. The column labelled 'excluded' gives the numbers of genes discarded from the analysis, based on their length (Methods 10). The 'included' column gives the remaining numbers of length-filtered SCP/TAPS genes in each species.



Supplementary Figure 32. Unique enzymes annotated and KEGG metabolic modules among the 81 nematode and platyhelminth species.

(a) Counts of enzymes (unique EC identifiers) in the 81 platyhelminth (red circle) and nematode (blue circle) species (i.e. present in any one of the set). (b) Distribution of conserved ECs across nematode and platyhelminth species (i.e. conserved across all species of the set). (c,d) Number of unique ECs annotated and KEGG modules deemed complete in all 81 species (in c) and the tier 1 species (those with high-quality assemblies; Methods 2) (in d), along with their proteome size. (e)

clustering of all 81 species based on the KEGG module presence (Jaccard similarity index and Ward's linkage). Bootstrap support values are indicated. Since hole-filling was only used to predict additional ECs for tier 1 species, ECs from hole-filling were not considered in this figure, except in part (d).

List of Supplementary Tables

Supplementary Table 1: Sample and sequencing information

Supplementary Table 2a: Properties of assemblies

Supplementary Table 2b: CEGMA genes missing from certain sets of species

Supplementary Table 2c: Origin of data used in comparison of gene sets in Compara

Supplementary Table 3: Gene set properties

Supplementary Table 4: Rationale, classification and taxonomy of species analysed

Supplementary Table 5: Repeat content

Supplementary Table 6: Results of Bayesian mixed effect model for genome size

Supplementary Table 7: GO annotation, predicted secretion data and transmembrane domains, and per-species gene counts for Compara gene families

Supplementary Table 8a: Counts of synapomorphic gene families at key nodes of the species tree

Supplementary Table 8b: Functional annotation of synapomorphic families

Supplementary Table 8c: Additional ferrochelatase protein sequences retrieved from NCBI

Supplementary Table 9a: 995 gene families selected based on variation scores

Supplementary Table 9b: Gene families excluded from expanded family analysis

Supplementary Table 10a: Counts of SCP/TAPS genes identified in the genomes

Supplementary Table 10b: Identifiers of SCP/TAPS genes identified in the genomes

Supplementary Table 11a: MEROPS protease family counts per species

Supplementary Table 11b: MEROPS protease inhibitor family counts per species

Supplementary Table 11c: Protease and protease inhibitor genes

Supplementary Table 12: Species life history traits

Supplementary Table 13: Parasite tissue tropism in definitive host

Supplementary Table 14a: Nematode-specific domain combinations

Supplementary Table 14b: Platyhelminth-specific domain combinations

Supplementary Table 15a: GPCR class, confidence level, and number of members in the final 230 Compara GPCR families

Supplementary Table 15b: Species-wise distribution of GPCRs in the final 230 Compara GPCR families

Supplementary Table 15c: Original putative GPCRs

Supplementary Table 15d: Final list of GPCRs

Supplementary Table 15e: Genes used as initial seeds to find GPCR Compara families

Supplementary Table 16a: Ligand gated ion channel predictions in our nematode and platyhelminth gene sets

Supplementary Table 16b: Accessions for ligand gated ion channels (from the WormBase ParaSite database) used in the phylogenetic analysis

Supplementary Table 16c: Accessions for the human and fruitfly ligand gated ion channel sequences used as outgroups in the phylogenetic analysis.

Supplementary Table 17: ABC transporter predictions

Supplementary Table 18a: Enzyme Commission number (EC) annotations for all 81 nematode and platyhelminth species

Supplementary Table 18b: Low confidence EC annotations based on membership of Compara families

Supplementary Table 18c: KEGG metabolic module completion for all 81 nematode and platyhelminth species, without using ECs from hole-filling

Supplementary Table 18d: KEGG metabolic module completion for the 33 tier 1 nematode and platyhelminth species, including ECs from hole-filling

Supplementary Table 18e: Heatmap of pathway conservation in KEGG pathways for nematodes and platyhelminths, without using ECs from hole-filling

Supplementary Table 18f: Heatmap of pathway conservation in KEGG pathways for nematodes and platyhelminths, using ECs from hole-filling

Supplementary Table 18g: Amino acid and vitamin auxotrophy predictions for tier 1 species, assigned by Pathway tools

Supplementary Table 18h: Metabolic chokepoints detected in helminths

Supplementary Table 19a: Beta-oxidation enzyme annotation

Supplementary Table 19b: Ketone bodies metabolism enzyme annotation

Supplementary Table 19c: Glyoxylate cycle specific enzyme annotation

Supplementary Table 19d: GABA-shunt enzyme annotation

Supplementary Table 19e: Glycine cleavage system enzyme annotation

Supplementary Table 19f: Pyrimidine synthesis enzyme annotation

Supplementary Table 19g: IMP biosynthesis enzyme annotation

Supplementary Table 19h: Haem synthesis enzyme annotation

Supplementary Table 20: KEGG pathways with most unique ECs (present in flatworms but absent in nematodes, or vice versa)

Supplementary Table 21a: Known anthelmintic drugs (for humans and veterinary animals), nematocides, and other anthelmintic compounds

Supplementary Table 21b: Known (or hypothesised) drug single-protein targets for known anthelmintic drugs and compounds

Supplementary Table 21c: Novel anthelmintic drug targets proposed in literature

Supplementary Table 21d: Potential anthelmintic drug targets

Supplementary Table 22: RNAseq used for gene-finding QC

Supplementary Table 23a: Kinase counts per kinase class and per Compara family

Supplementary Table 23b: Kinase gene identifiers in nematode and platyhelminth species

Supplementary Table 24a: InterPro (IPR) enrichment in different species groups

Supplementary Table 24b: GO enrichment in different species groups

Supplementary Table 24c: Pfam enrichment in different species groups

Supplementary Table 25: CAZYME predictions for the 81 nematodes and platyhelminths