*Nat Genet*. Author manuscript; available in PMC 2016 September 06.

Published in final edited form as:

*Nat Genet*. 2014 July ; 46(7): 693–700. doi:10.1038/ng.3010.

# Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction

**Bernardo J. Foth**[#1], **Isheng J. Tsai**[#1,2], **Adam J. Reid**[#1], **Allison J. Bancroft**[#3], **Sarah Nichol**[1], **Alan Tracey**[1], **Nancy Holroyd**[1], **James A. Cotton**[1], **Eleanor J. Stanley**[1], **Magdalena Zarowiecki**[1], **Jimmy Z. Liu**[4], **Thomas Huckvale**[1], **Philip J. Cooper**[5,6], **Richard K. Grencis**[3], and **Matthew Berriman**[1]

[1]Parasite Genomics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

[2]Division of Parasitology, Department of Infectious Disease, Faculty of Medicine, University of Miyazaki, Miyazaki 889-1692, Japan

[3]Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester M13 9PT, UK

[4]Statistical Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

[5]Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK

[6]Centro de Investigación en Enfermedades Infecciosas, Escuela de Biología, Pontificia Universidad Católica del Ecuador, Quito, Ecuador

[#] These authors contributed equally to this work.

## Abstract

Whipworms are common soil-transmitted helminths that cause debilitating chronic infections in man. These nematodes are only distantly related to *Caenorhabditis elegans* and have evolved to



Correspondence and requests for materials should be addressed to M.B. (mb4@sanger.ac.uk) and R.K.G. (richard.grencis@manchester.ac.uk).

**Author Contributions**: AJB cultivated *T. muris* parasites in mice and prepared DNA and RNA. PJC provided *T. trichiura* parasite material. JIT assembled the genomes and produced chromosomal assignments. AT and SN manually improved the genome assemblies and curated genes. BJF, EJS, and JIT trained genefinding software and structurally annotated the genomes. BJF and MZ provided functional gene annotation and analysis. JAC performed gene family clustering and analysis. BJF analysed the *T. muris* transcriptome, genome read coverage and heterozygosity. AJR analysed the mouse transcriptome. MB and MZ analysed drug targets. NH coordinated the project and managed the sequencing. JZL carried out the GWAS analysis. AJB, AJR, BJF, MB, and RKG wrote the manuscript. MB and RKG directed the project.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

**Author Information**

The *Trichuris* genome projects have been registered under the INSDC project ID numbers PRJEB2108 (*T. muris*) and PRJEB2679 (*T. trichiura*). Genome data and the complete genome annotation are available for download from ftp://ftp.sanger.ac.uk/pub/pathogens/Trichuris/GenomePaper2014. Genome sequencing data have been deposited in the European Nucleotide Archive (ENA) under sample accession numbers ERS016744, ERS016965, ERS056020, and ERS244696 for *T. muris*, and under ERS056020 for *T. trichiura*. RNA-seq transcriptome data have been deposited in the ENA under study accession numbers ERP002000 (*T. muris*) and ERP002560 (mouse) and in ArrayExpress under study accession numbers E-ERAD-125 (*T. muris*) E-ERAD-181 (mouse).

occupy an unusual niche, tunneling through epithelial cells of the large intestine. Here we present the genome sequences of the human-infective *Trichuris trichiura* and the murine laboratory model *T. muris*. Based on whole transcriptome analyses we identify many genes that are expressed in a gender- or life stage-specific manner and characterise the transcriptional landscape of a morphological region with unique biological adaptations, namely bacillary band and stichosome, found only in whipworms and related parasites. Using RNAseq data from whipworm-infected mice we describe the regulated Th1-like immune response of the chronically infected cecum in unprecedented detail. *In silico* screening identifies numerous potential new drug targets against trichuriasis. Together, these genomes and associated functional data elucidate key aspects of the molecular host-parasite interactions that define chronic whipworm infection.

Whipworms (*Trichuris trichiura*) infect an estimated 700 million people worldwide[1]. They are one of three major groups of soil-transmitted helminths (STH), the others being *Ascaris* and hookworms, that impede socioeconomic development of entire populations. *Trichuris* parasites invade the epithelium of the colon of their hosts, particularly the cecum, where they may persist for years and can cause colitis, anaemia, and Trichuris dysentery syndrome. *T. muris* is an established murine laboratory model for human trichuriasis and shares with *T. trichiura* many aspects of its biology including the specialized niche within the host. Work on *T. muris* has been pivotal in defining the crucial role of Th2 immune responses and of IL-13 in protective immunity to parasitic helminths[2], while deliberate infection with whipworm eggs, typically of the porcine parasite *T. suis*, is emerging as a novel therapy option for patients with immune-mediated diseases[3,4]. Whipworms have a simple and direct life cycle, and unlike the related parasite *Trichinella spiralis,* whipworm larvae do not form cysts in muscle tissue but reside exclusively in the intestine. Bacillary cells and stichocytes are distinctive cell types found only in clade I nematodes[5] and are located in the slender anterior part of the adult whipworm that is burrowed within the intestinal epithelium. The bulbous posterior end of the whipworm lies free in the intestinal lumen and harbours the reproductive organs, giving adult *Trichuris* parasites their characteristic whip-like morphology.

Here we present high quality genome sequences for *T. trichiura* and *T. muris*, the first duo of a major human STH and its murine counterpart. We resolve chromosomal sequences and infer sex chromosome-specific parasite genes and new potential drug targets. Based on high-throughput transcriptomics data we identify whipworm proteins that are highly expressed in the anterior region of the parasite that is in intimate contact with the cytoplasm of host intestinal cells and with the immune system. Gene expression data from mice with low-dose whipworm infection provide a detailed description of a regulated Th1-like immune response to the infected cecum that is not limited to the immediate site of infection. The availability of these two important whipworm genomes and the integration of parasite and host data presented here will underpin future efforts to control these parasites and exploit their immunological interplay for human benefit.




Europe PMC Funders Author Manuscripts

# RESULTS

## Genome structure and content

We produced a 75.2 megabase (Mb) high quality draft genome assembly from a clinically isolated adult male *T. trichiura* using Illumina technology. In parallel, multiple technologies and manual finishing were used to produce an 85 Mb reference genome assembly from the more readily available rodent model species *T. muris* (Supplementary Methods). More than half of the *T. muris* and *T. trichiura* genomes assembled into scaffolds of at least 1,580 kilobases (kb) and 71 kb, respectively (Table 1). Despite their differences in contiguity both genomes were measured as 94.8% complete using the Core Eukaryotic Genes Mapping Approach (CEGMA)[6]. The *T. muris* genome comprises two pairs of autosomes and one pair of X and Y sex chromosomes (2n = 6)[7], while the karyotype of *T. trichiura* has so far not been reported. In *Trichinella* species of the same phylogenetic clade, the XO-sex determination system is used with 2n = 6 for female (XX) and 2n = 5 for male (X0) worms[8]. To explore the architecture of the genome, assembly scaffolds of *T. muris* were joined into super-scaffolds by optical mapping (assembly version 4.0, Table 1). Clustering of one-to-one ortholog patterns between *T. muris* and *Trichinella spiralis* revealed three distinct linkage groups in each species, providing strong evidence that chromosome-level synteny has been preserved across genera (Fig. 1a). Contrasting with the observed macro-synteny, small-scale gene order has been mostly lost between all three species (Fig. 1a and Supplementary Fig. 1). Such significant intra-chromosomal rearrangements have also been found in other nematode lineages[9-12] and appear to be a hallmark of nematode genome evolution.

To assign genomic scaffolds to chromosomal locations, we used mapped resequencing data from females (*T. muris* only) and single males (*T. muris* and *T. trichiura*). From the males, one linkage group could be putatively identified as the X chromosome, based on an almost complete lack of heterozygosity and half the relative sequence coverage compared with the other two linkage groups that are presumed to be pairs of autosomes (Fig. 1b,c, Supplementary Fig. 2a-c, Supplementary Table 1). In contrast, the putative X-chromosomal scaffolds from *T. muris* females showed equivalent read coverage and a similar level of heterozygosity to the putative autosomes. Furthermore, the putative X-chromosomal scaffolds are significantly enriched for genes that are transcriptionally upregulated in female worms (Supplementary Table 2.1).

From differential sequence coverage in males and females (Supplementary Fig. 2e-f, Supplementary Table 1.3, Supplementary Methods), putative *T. muris* Y chromosome sequences were identified equivalent to 24.4 Mb, when the estimated true copy numbers of the significant number of collapsed repeats were taken into account (Supplementary Table 1.4). Despite their similar estimated sizes, the *T. muris* X chromosome (26.9 Mb) contains 2,137 genes whereas the Y chromosome has a non-redundant gene content of less than 70 genes, of which most are of unknown function or transposon-related (Supplementary Table 2.2). We further identifed putative centromeric repeat sequences on three *T. muris* scaffolds that together are estimated to constitute 5.33 Mb (Supplementary Table 1.4, Supplementary Fig. 3, Supplementary Methods).

In *T. trichiura*, 9,650 genes were predicted and in *T. muris*, aided by deep transcriptome sequencing, 11,004 genes were found (Table 1). Predicted proteins from both species were clustered together with sequences from other nematodes and outgroups into a total of 7,354 gene families. Of these, 5,868 families contained genes from both species, i.e. 6,629 and 6,641 genes of *T. muris* and *T. trichiura*, respectively. Thus, the majority of *Trichuris* genes are shared orthologs between both species (Fig. 2a,b). Despite the widespread loss of gene colinearity, the utility of the rodent model is emphasised by their highly similar proteomes; 5,060 one-to-one protein orthologs exhibit an average amino acid identity of 79%. 2,350 *T. trichiura* and 3,817 *T. muris* genes appear to be species-specific within this comparison and are particularly enriched in hypothetical proteins of no known function (e.g. 76% of the *T. muris* species-specific genes vs 35% genome-wide). Of the altogether 3,953 and 2,014 hypothetical proteins in *T. muris* and *T. trichiura*, respectively, 434 (11.0%, *T. muris*) and 253 sequences (12.6%, *T. trichiura*) are predicted by Phobius to contain signal peptides and 517 (13.1%, *T. muris*) and 359 sequences (17.8%, *T. trichiura*) to contain transmembrane domains. For those that have functional annotation, Gene Ontology term enrichment analysis suggests extracellular proteins, proteases and protease inhibitors are particularly abundant amongst the species-specific proteins in *T. muris*. One large gene family of unknown function in *T. muris* consists of a lineage-specific expansion of over 40 *T. muris* genes and 6 genes in *T. trichiura* but only one or two copies in other nematode species (FAM217; Supplementary Table 3, and Supplementary Data file 1).

## Patterns of parasite gene expression

To better understand the parasite's unusual niche within the mucosa of the large intestine we generated and compared transcriptome-wide gene expression data for several biological samples of *T. muris*, followed by protein domain and GO-term enrichment analysis. The stichosome and bacillary band in the whipworm anterior have been suggested to be involved in nutrient uptake, digestion and host-parasite interactions, for which they would be perfectly placed morphologically. Nevertheless, their exact biological roles are still poorly understood. Our data show that transcripts in the *T. muris* anterior region are dominated by chymotrypsin A-like serine proteases and by protease inhibitors with similarity to secretory leukocyte peptidase inhibitor (SLPI) (MEROPS families S1A and I17, respectively; Supplementary Tables 4, 5; Supplementary Fig. 4). Chymotrypsin A-like serine proteases are encoded by over 75 genes in each of the two *Trichuris* genomes, which is far more than in other nematodes, making them the single largest group of proteases (Supplementary Table 4). In *T. muris*, three quarters of these (63) are transcriptionally upregulated in the anterior region, and of these two thirds (42) are likely secreted (Supplementary Tables 6, 7, Supplementary Fig. 4). These proteases may therefore serve digestive or host-immunomodulatory functions or degrade host intestinal mucins that act as a physical barrier to the parasite[13,14]. In *Trichuris*, serine proteases are thus more abundant both in terms of gene number and gene expression than other protease families (in particular cysteine proteases), which is in contrast to the situation in many other nematodes[15-17] (Supplementary Tables 4, 5).

Eighty of the 111 protease inhibitors found in *T. muris* encode serine protease inihibitors (MEROPS families I1, I2, I15, and I17), with SLPI-like proteins (i.e. proteins comprised

mostly of WAP domains) collectively being by far the most abundant protease inhibitor transcripts in the anterior region (Fig. 3a). WAP domains are found in a number of proteins that also include the mammalian Whey Acidic Protein (WAP) and elafins. The *T. muris* genome contains 44 genes with between one and 9 WAP domains, while *T. trichiura* and *Trichinella spiralis* feature 20 and 23 such proteins, respectively. Surprisingly, all three trichocephalid species each contain only a single WAP domain-containing protein (TMUE_s0077003100, TTRE_0000351901, EFV57447) in which the WAP domain contains the canonical 8 cysteine residues that have given this domain its alternative name '4-disulphide core'. These proteins are large and bear some similarity to the mesocentin protein of *C. elegans* which exhibits RNAi phenotypes such as neuroanatomical defects affecting chemosensory neurons and axons[18]. In contrast, all other WAP domains of *Trichuris* and *Trichinella* have an apparently novel trichocephalid adaptation of exactly six cysteine residues (Fig. 3b and Supplementary Fig. 5). How this modification affects WAP domain protein function is unknown. In addition to their protease inhibitor activity, both mammalian SLPIs and elafins have immunomodulatory, anti-inflammatory and antimicrobial properties with a role in innate immune defence and wound healing[19-21]. Epithelial cells frequently produce these proteins in response to inflammation, to modulate the inflammation process, cytokine secretion, and cell recruitment, and to favour a Th1-type immune response[19]. The fact that there are apparently no SLPI-like proteins present in other parasitic helminth lineages, that they are strongly and specifically expressed in the whipworm anterior region, and that the majority of SLPI-like proteins are likely secreted suggests that this gene family carries out nematode clade I-specific functions that may include the inhibition of inflammation in the host intestinal epithelial tissue as it is being invaded and wounded by the parasite.

DNase II-like proteins are also particularly highly expressed in the whipworm anterior region (Fig. 4a). *Trichinella spiralis* contains 166 DNase II-like proteins[22,23] that include the abundant excretory-secretory protein gp43[24]. *T. muris* and *Trichinella spiralis* each encode one ortholog of the metazoan DNase II (Fig. 4b and Supplementary Fig. 6), which for *T. muris* exhibits low and uniform transcript levels across the biological samples analysed. In contrast, all other trichocephalid DNase II-like proteins are so divergent that they are positioned outside the major well-supported cluster in a phylogenetic tree (Fig. 4b). As in other animals[25], the three DNase II-like nucleases in *C. elegans* (NUC-1, CRN-6, and CRN-7) are involved in apoptosis and development[26]. A homologous protein of starfish (plancitoxin) has been reported to be able to enter the nuclei of rat liver epithelial cells and cause caspase-independent apoptosis[27]. One intriguing potential biological role of DNase II-like proteins in trichocephalid parasites is the degradation of host DNA that may be released when the parasites damage host tissue during invasion, thereby limiting inflammatory and immune responses that could be elicited by undigested host DNA[22].

We find male-specific expression (Supplementary Table 8) for proteins with Major Sperm Protein (MSP) domains, which likely play a role in the amoeboid locomotion of nematode sperm[28], and for proteins with casein kinase-related and EGF-like domains which may be involved in male mating-associated functions[29]. In contrast, proteins with chitin-binding domains are upregulated in female whipworms. As nematode eggshells commonly contain chitin, it is likely that these proteins are associated with the formation of the eggshell or a

related function[30]. The transcriptional landscape of L2 and L3 larvae is similar to that of the anterior region of adult worms, which is likely due both to the shared intraepithelial location and the shared absence of reproduction- and gender-specific transcripts that dominate the worm posterior (Supplementary Table 8). In addition, larval stages show high expression of ribosomal and of collagen- and fibronectin-related proteins, which likely reflects the fast growth and associated protein and cuticle synthesis in the larval stage whipworms (Supplementary Fig. 4, Supplementary Table 8, 9).

## Host intestinal response to infection

A low dose infection of mice with *T. muris* leads to chronic infection whereas high dose infections are typically cleared in the majority of inbred mouse strains. Therefore, low dose infection of C57BL/6 mice with *T. muris* is used as a model of natural human *T. trichiura* chronic infection which usually presents with a low parasite burden[31] and may also provide a model for inflammatory immune diseases of the gut[32,33]. A Th2 response is required for resolution of infection and acquisition of immunity[34], but chronic infections are typified by a regulated Th1 response[35,36]. This combined response favours the parasite, counteracting a protective Th2 response, resulting in slower epithelial turnover, leading to pathology with features of colitis but preventing severe intestinal pathology[37]. To investigate the precise nature of these responses, we characterized gene expression changes in mouse cecum and mesenteric lymph node (MLN) upon infection, using RNA-seq. The changes in host gene expression were common between the precise site of worm occupation and worm-free areas. We found only five genes differentially expressed between these tissues in infected animals, suggesting broad regulatory activity within the cecum.

Within infected cecum, we found 868 genes upregulated and 590 downregulated, compared to the same tissue in naïve mice. Upregulated genes were enriched for a variety of functional terms associated with the immune system suggesting this is the primary mediator of interaction with the worm during chronic infection. As expected, these changes were consistent with a Th1 response: CD4 was upregulated 4-fold, while pro-inflammatory cytokines IFN-gamma and TNF-alpha were upregulated 26 and 12-fold respectively (Supplementary Fig. 7). Conversely IL-4, IL-5, IL-9 and IL-13, typical of a Th2 response, were not differentially expressed. IL-18 has been suggested to have a role in promoting susceptibility by inhibiting the Th2 response[38], however at this stage of infection we saw a downregulation of IL-18 in susceptible mice suggesting instead a role in downregulating an excessive Type 1 response. We observed upregulation of IL-16 and CCR5, which have been implicated in Th1 cell migration[39], but have not previously been associated with *Trichuris* infection. There was strong upregulation of immunoglobulin classes G2B, G2C, A and M, but not E in the cecum post-infection reflecting the changes seen in the MLN (Supplementary Fig. 7, Supplementary Table 10). This is likely a result of IFN-$\gamma$ stimulation but a role for these isotypes in chronic infection is unknown. Certainly, antibody *per se* is thought to be dispensable for protection against primary infections[40]. In chronic infections, the antibody response seen here may be a non-functional by-product of a parasite-induced Th1 response or instead a response to bacterial infection associated with parasite-induced intestinal damage. Support for the latter comes from the observed secretory IgA response, which is commonly associated with commensal bacteria[41]. Supplementary Table 11 shows

the full extent of cytokine and chemokine differential expression in infected cecum. In the MLN, 381 genes were upregulated while 176 were downregulated after infection. The gene set was dominated by genes related to the cell cycle and IgG, suggesting prolonged production of B cells. Whether this simply indicates response to chronic intestinal infection, increased exposure to microflora or that these B cells or isotypes have specific or novel functional roles, remains to be determined.

In a chronic infection it is essential that the parasite limits damage to the host. In particular IL-10 has been proposed to control tissue repair during chronic infection[35]. Furthermore IL-22, another member of the IL-10 superfamily, has been implicated in therapeutic *T. trichiura* infection[42]. We observe IL-10 and IL-22 to be upregulated in infected cecum. The source of these cytokines could be the classicly induced regulatory FoxP3+ Treg cells, however we do not observe upregulation of associated markers FoxP3 and CD25 suggesting that the source may be other IL-10 producing T cell subsets. Mucosal mast cells have previously been observed in acute infection with *T. muris*, but do not appear to be responsible for parasite resistance[43]. Evidence from the AKR mouse model of chronic infection[44] and here from the low dose C57BL/6 model suggest that they increase in number during chronic infection and therefore may have a role in tissue repair or immunoregulation perhaps by the production of IL-10 (Supplementary Fig. 7). We observe several markers of M2 (alternatively activated) macrophages (Supplementary Fig. 7), which have also been shown to be dispensable for resistance to *Trichuris* infection[45]. They are thought to be specific to a Th2 response and although we do not detect upregulation of IL-4 or IL-13, very low levels of these could be present as we detect no reads for their transcripts in uninfected samples and low levels of reads in infected ones. M2 macrophages are known to have roles in tissue repair[46] and their dispensability for parasite clearance and presence in chronic infection indicates that this may be their function in *Trichuris* infection.

The pathogenesis caused by *T. muris* infection of C57BL/6 mice is similar to that seen in human Ulcerative Colitis (UC). Furthermore by crossing mice resistant and susceptible to *T. muris* infection, QTLs have been identified which are shared with other experimental models of colitis[32]. Here we show that *T. muris* infection causes changes in expression of a significant number of genes associated through genome-wide association studies with UC and several other inflammatory diseases (Supplementary Table 12, Supplementary Fig. 8). This provides support for the use of *T. muris* as a model of these devastating diseases prevalent in countries relatively free of intestinal parasites in addition to its role as a model of *T. trichiura* infection.

## Insights into new drug targets

Recent meta-analyses have revealed that albendazole and mebendazole, used to treat trichuriasis, may only completely clear about one third of infections, and high rates of re-infection are commonly encountered[47,48]. With the unsatisfactory performance of current anthelmintics and the availability of complete genome sequences, new opportunities for target-based approaches to drug discovery need to be explored. The preceding sections of this study highlight parasite processes that could potentially be disrupted by drugs (e.g. Proteases, protease inhibitors, DNAseII) but by combining our transcriptome data with

results of database searches, we have assigned, on a genome wide scale, potentially desirable properties for drug target selection. These properties can be weighted and ranked (Supplementary Tables 13, 14) or simply filtered. For instance, 8,307 genes are expressed in adult whipworms, 4,269 genes have high inferred druggability, and for 600 we inferred evidence of essentiality. Fulfilling all of these criteria, we find only 397 putative targets. This list was further reduced to 29 by filtering for candidate proteins with homologs that are targets of existing approved drugs (Table 2, Supplementary Tables 13, 14). Amongst these, we identify fatty acid synthase, the target of the anti-obesity drug orlistat; DNA topoisomerase, a target for the previously utilised anti-schistosomiasis drug lucanthone; and calmodulin, the target for the anti-diarrheal drug loperamide, which has previously been reported to increase the efficacy of mebendazole in treating trichuriasis[49].

## DISCUSSION

Whipworms are exquisitely well adapted to their unusual biological niche. They are able to invade and maintain over extended periods of time their position within the epithelium of the large intestine, one of the fastest renewing tissues of the body, and often without causing excessive pathology. The anterior end of adult worms appears to be key for host-parasite interactions, immunomodulation, and feeding. This is because it is in intimate contact with host tissue and because it is home to two specialised cellular structures: the bacillary band and the stichosome. These adaptations are unique to whipworms and related parasites, i.e. organisms that spend part of their life cycle buried in the intestinal mucosa. This study identifies numerous genes with anterior end-specific expression that include protease inhibitors, DNase II, and serine proteases. Many of these molecules are likely secreted from the worm and are plausibly involved in both immunomodulation and feeding. How and where secreted proteins are actually released from adult worms remains a fundamental and unanswered question; they could either be routed via the esophagus and exit via the anus into the intestinal lumen or find their way to the exterior of the worm via the intriguing pores of the bacillary cells that are in direct contact with host cell cytoplasm[50]. It is also unknown whether feeding actually occurs via the mouth, as *Trichuris* lacks the muscular pharynx of other nematodes that pump food through the gut. Alternatively, whipworms could both secrete digestive enzymes (together with immunomodulatory molecules) and absorb nutritional molecules via the pores of the bacillary cells.

Here, we also present the landscape of transcriptomic changes in chronically infected host tissue and suggest aspects of the host immune system that might be involved in reducing pathology. Whether this is mediated by alternatively activated macrophages, mast cells or other cell populations, determining how the interplay of host and parasite regulates the immune system will also help us to better understand diseases such as ulcerative colitis that appear to have much in common with whipworm infection. Intriguingly, the pig whipworm *T. suis* is licensed as a medicine for treating colitis. Comparative immunology of these infections should help us to develop more targeted therapies in the future. Finally, this work emphasises the power of combining genomics and transcriptomics to study an important human pathogen and its well-defined model system in tandem, in not only opening up novel avenues of future therapeutics in human disease but also providing fundamental biological data towards our understanding of the host parasite relationship.

# ONLINE METHODS

Methods and materials are described here in brief. For further details please see the Supplementary Note. All animal experiments were performed under the auspices of the University of Manchester ethical review committee and under the Home Office Scientific Procedures Act (1986).

## Genome sequencing

The whipworm genome sequences are based on a single male individual of *T. trichiura* from Ecuador and a pool of *T. muris* parasites Edinburgh strain. For *T. trichiura*, the study protocol was approved by the ethics committees of the Hospital Pedro Vicente Maldonado, Pichincha Province, and Pontificia Universidad Catolica del Ecuador, Quito, Ecuador, and included appropriate informed written consent. Genome sequencing was performed using both Illumina (*T. trichiura* and *T. muris*) and 454 (*T. muris*) platforms. To generate separate male and female *T. muris* samples (one male individual and a pool of eleven females), worms were separated by gender based on size and morphology. The sequenced libraries are listed in Supplementary Table 15.

## Genome assembly and improvement

Genome assembly was carried out with SGA[52] v0.9.17 and Velvet[53] v1.2.03 (*T. trichiura*) as well as Celera[54] v7.0 (*T. muris*). Misassemblies were identified and corrected using REAPR[55] and manual inspection. Genome assemblies were scaffolded and improved using SSPACE[56], IMAGE[57], Gapfiller[58], and (for *T. muris*) by incorporating optical map data with tools from OpGen. For *T. muris*, genome assembly v2.1 was used to predict genes, whereas assembly v4 represents the most recent assembly with the best long-range contiguation for which sequence scaffolds were joined also in cases where the corresponding gaps were too large to be spanned by sequencing reads (i.e. larger than ~10kb) but were confidently spanned by optical map contigs.

## Transcriptome sequencing - *T. muris* and mouse

High-throughput transcriptome data were generated from RNA of *T. muris* and mouse tissue (C57BL/6). For larval stage and adult whipworms, RNA was prepared using TRIZOL and lysing matrix D (1.4 mm ceramic spheres) and a Fastprep24 (MP Biomedicals). Mice had been subjected to a low-dose infection with *T. muris* (25 eggs by oral gavage) or were uninfected, and the sampled tissues included mesenteric lymph node, a section of cecum where the worms reside and - as a control - an uninfected section of cecum. RNAseq libraries were prepared following the RNAseq protocols of the Illumina mRNA-Seq Sample Prep Kit and the Illumina TruSeq kit. The transcriptome libraries were sequenced on Illumina HiSeq 2000 machines and are listed in Supplementary Table 16.

## Gene predictions

For *T. muris*, the gene predictor Augustus[59] v2.4 was trained on a gene training set - derived from CEGMA[6] predictions, Augustus predictions, and manual curation - and was run by also providing intron hints based on RNAseq data. Genes of particular interest and genes that appeared incorrect based on semi-automatic screens and manual inspection were

manually curated. Likely transposon-related genes were removed from the final gene set. For *T. trichiura*, genes were predicted using a Maker[60] v2.2.28 pipeline that incorporated *ab initio* predictions by Augustus, GeneMark-ES[61] v2.3a (self-trained) and SNAP[62] 2013-02-16 and considered gene models based on comparison with other species.

### Functional gene annotation

Functional gene annotation including the assignment of gene product descriptions and GO terms was based on Interpro protein domain analysis and BLAST searches against annotated genomes. Pfam protein domain predictions and GO term assignments as provided by Interproscan are listed in Supplementary Tables 17 and 18. Genes lacking both Pfam domain and BLAST hit were labelled “hypothetical protein”. For *T. trichiura*, genes with a one-to-one ortholog in *T. muris*, functional annotation was transferred from the *T. muris* ortholog.

### Gene family clustering and phylogenetic analysis

Gene families were determined using orthoMCL[63] v2.0, and one-to-one orthologs with Inparanoid[64] v4.1 and OMA[65] v0.99t. Multiple sequence alignments were created using MAFFT[66] v6.857 and GBlocks[67] v0.91b, and phylogenetic trees were constructed with RAxMLHPC[68] v7.2.8. Sequences to be included in the phylogenetic tree for DNase II were selected based on the presence of the corresponding Interpro protein domain IPR004947.

### Chromosome-level analysis

The assignment of scaffolds and genes to chromosomes was performed by first clustering the numbers of one-to-one orthologs between the largest genome assembly scaffolds of *T. muris* and *Trichinella spiralis* which yielded three distinct “linkage groups”. More scaffolds of *T. muris* could then be assigned to the linkage groups based on the presence of gene orthologs in comparison to the large *T. spiralis* scaffolds that were already assigned to a linkage group. Sequence read coverage was determined by aligning Illumina data against the relevant genome assembly using SMALT v0.7.4 followed by running the command genomecov of BEDTools[69] v2.17.0. Heterozygosity per 10kb window of genomic sequence was calculated based on a pileup including base and variant calls that was generated by SAMtools[70] mpileup v0.1.18 from the high throughput sequence read alignment. Genomic scaffolds representing the X chromosome were identified based on both read coverage and heterozygosity. Mean read coverage over parts of the genome assembly with extraordinarily high read coverage was also used to estimate the real extent of DNA content in the parasite represented by such parts of the assembly. Centromeric sequences were putatively identified based on high read coverage and the length of the underlying repeat units (164 bp to 176 bp).

### Gene expression analysis - *T. muris* and mouse

For differential gene expression analysis, paired-end Illumina RNAseq data were mapped using Tophat[71] v1.4.1 to either the *T. muris* genome v2.1 (for *T. muris* samples) or a combined reference of mouse (mm10) and *T. muris* transcripts (for mouse samples). The number of reads per gene were enumerated with BEDTools (*T. muris*) or eXpress[72] (mouse), and differential expression analysis was carried out using the Bioconductor packages

edgeR[73] v3.2.4 (*T. muris*) and DESeq[74] (mouse). GO term enrichment analysis was carried with TopGO[75] and innateDB[76], and protein domain enrichment analysis was based on the results of Interproscan[77].

### Identification of novel drug targets

The ranking of all *T. muris* proteins for their suitability as a drug target was based on a number of different types of information: *T. muris* RNAseq expression data; orthology (as determined by OMA) of *T. muris* protein sequences to those in *T. trichiura, C. elegans*, mouse, human and drug targets; protein homolog essentiality in mouse and *C. elegans*; whether a protein is predicted to be an enzyme, based on KEGG orthology; druggability information from ChEMBL; and drug target information from DrugBank and from TTD (Therapeutic Targets Database). Nutraceutical targets were filtered out.

### GWAS analysis

A test to identify genes that are both differentially expressed in the cecum of whipworm-infected vs. uninfected mice and that are associated with immune-mediated diseases was performed as follows. The mouse genes were filtered for those that have a unique human ortholog, are annotated as protein-coding, and are located on autosomes. Lists of associated loci from published genome-wide association studies (GWAS) were extracted for four immune-mediated complex diseases (Crohn's disease, ulcerative colitis, celiac disease, and type 1 diabetes), as well as two likely immune-unrelated complex traits (height and body mass index). Testing for enrichment was performed using a Monte Carlo simulation approach that accounts for linkage disequilibrium between associated SNPs and non-random arrangement of functionally related genes within the genome.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

This work was funded by the Wellcome Trust through their support of the Wellcome Trust Sanger Institute (grant 098051) and an Investigator Award (WT100290MA) and a Programme grant (WT083620MA) to RKG. The collection of samples in Ecuador was supported by the Wellcome Trust (grant 088862/Z/09/Z). We thank Rick Maizels for initially proposing a *Trichuris muris* genome project, Carl Anderson and Tim Raine for sharing their unpublished method for enrichment analysis of IBD-associated genes, Seona Thompson for preparing parasite material, Taisei Kikuchi and Zahra Hance for preparing genomic DNA, Hayley M. Bennett for preparing RNA, Karen L. Brooks and Helen Beasley for manually curating gene models, Matt Dunn for generating optical map data, Neil Rawlings for helpful discussions, Avril Coghlan for supporting bioinformatic analysis, and Martin Aslett for data submission to EMBL.

## References

1. Bethony J, et al. Soil-transmitted helminth infections: ascariasis, trichuriasis, and hookworm. Lancet. 2006; 367:1521–32. [PubMed: 16679166]
2. Grencis RK, Bancroft AJ. Interleukin-13: a key mediator in resistance to gastrointestinal-dwelling nematode parasites. Clin Rev Allergy Immunol. 2004; 26:51–60. [PubMed: 14755075]
3. Summers RW, et al. Trichuris suis seems to be safe and possibly effective in the treatment of inflammatory bowel disease. Am J Gastroenterol. 2003; 98:2034–41. [PubMed: 14499784]

Europe PMC Funders Author Manuscripts


4. Jouvin MH, Kinet JP. Trichuris suis ova: testing a helminth-based therapy as an extension of the hygiene hypothesis. J Allergy Clin Immunol. 2012; 130:3–10. quiz 11-2. [PubMed: 22742834]
5. Parkinson J, et al. A transcriptomic analysis of the phylum Nematoda. Nat Genet. 2004; 36:1259–67. [PubMed: 15543149]
6. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007; 23:1061–7. [PubMed: 17332020]
7. Spakulova M, Kralova I, Cutillas C. Studies on the karyotype and gametogenesis in Trichuris muris. J Helminthol. 1994; 68:67–72. [PubMed: 8006389]
8. Mutafova T, Dimitrova Y, Komandarev S. The karyotype of four Trichinella species. Z Parasitenkd. 1982; 67:115–20. [PubMed: 7072318]
9. Desjardins CA, et al. Genomics of Loa loa, a Wolbachia-free filarial parasite of humans. Nat Genet. 2013; 45:495–500. [PubMed: 23525074]
10. Kikuchi T, et al. Genomic insights into the origin of parasitism in the emerging plant pathogen Bursaphelenchus xylophilus. PLoS Pathog. 2011; 7:e1002219. [PubMed: 21909270]
11. Mitreva M, et al. The draft genome of the parasitic nematode Trichinella spiralis. Nat Genet. 2011; 43:228–35. [PubMed: 21336279]
12. Ghedin E, et al. Draft genome of the filarial nematode parasite Brugia malayi. Science. 2007; 317:1756–60. [PubMed: 17885136]
13. Hasnain SZ, McGuckin MA, Grencis RK, Thornton DJ. Serine protease(s) secreted by the nematode Trichuris muris degrade the mucus barrier. PLoS Negl Trop Dis. 2012; 6:e1856. [PubMed: 23071854]
14. Drake LJ, Bianco AE, Bundy DA, Ashall F. Characterization of peptidases of adult Trichuris muris. Parasitology. 1994; 109(Pt 5):623–30. [PubMed: 7831097]
15. Marcilla A, et al. The transcriptome analysis of Strongyloides stercoralis L3i larvae reveals targets for intervention in a neglected disease. PLoS Negl Trop Dis. 2012; 6:e1513. [PubMed: 22389732]
16. Cantacessi C, et al. Massively parallel sequencing and analysis of the Necator americanus transcriptome. PLoS Negl Trop Dis. 2010; 4:e684. [PubMed: 20485481]
17. Cantacessi C, et al. The transcriptome of Trichuris suis--first molecular insights into a parasite with curative properties for key immune diseases of humans. PLoS One. 2011; 6:e23590. [PubMed: 21887281]
18. Benard CY, Boyanov A, Hall DH, Hobert O. DIG-1, a novel giant protein, non-autonomously mediates maintenance of nervous system architecture. Development. 2006; 133:3329–40. [PubMed: 16887823]
19. Williams SE, Brown TI, Roghanian A, Sallenave JM. SLPI and elafin: one glove, many fingers. Clin Sci (Lond). 2006; 110:21–35. [PubMed: 16336202]
20. Scott A, Weldon S, Taggart CC. SLPI and elafin: multifunctional antiproteases of the WFDC family. Biochem Soc Trans. 2011; 39:1437–40. [PubMed: 21936829]
21. Wilkinson TS, Roghanian A, Simpson AJ, Sallenave JM. WAP domain proteins as modulators of mucosal immunity. Biochem Soc Trans. 2011; 39:1409–15. [PubMed: 21936824]
22. Liu MF, et al. The functions of Deoxyribonuclease II in immunity and development. DNA Cell Biol. 2008; 27:223–8. [PubMed: 18419230]
23. Mitreva M, Jasmer DP. Advances in the sequencing of the genome of the adenophorean nematode Trichinella spiralis. Parasitology. 2008; 135:869–80. [PubMed: 18598573]
24. Bien J, et al. Comparative analysis of excretory-secretory antigens of Trichinella spiralis and Trichinella britovi muscle larvae by two-dimensional difference gel electrophoresis and immunoblotting. Proteome Sci. 2012; 10:10. [PubMed: 22325190]
25. Crow YJ. The story of DNase II: a stifled death-wish leads to self-harm. Eur J Immunol. 2010; 40:2376–8. [PubMed: 20706989]
26. Lai HJ, Lo SJ, Kage-Nakadai E, Mitani S, Xue D. The roles and acting mechanism of Caenorhabditis elegans DNase II genes in apoptotic dna degradation and development. PLoS One. 2009; 4:e7348. [PubMed: 19809494]

27. Ota E, et al. Caspase-independent apoptosis induced in rat liver cells by plancitoxin I, the major lethal factor from the crown-of-thorns starfish Acanthaster planci venom. Toxicon. 2006; 48:1002–10. [PubMed: 16973201]
28. Tarr DE, Scott AL. MSP domain proteins. Trends Parasitol. 2005; 21:224–31. [PubMed: 15837611]
29. Hu J, Bae YK, Knobel KM, Barr MM. Casein kinase II and calcineurin modulate TRPP function and ciliary localization. Mol Biol Cell. 2006; 17:2200–11. [PubMed: 16481400]
30. Johnston WL, Krizus A, Dennis JW. Eggshell chitin and chitin-interacting proteins prevent polyspermy in C. elegans. Curr Biol. 2010; 20:1932–7. [PubMed: 20971008]
31. Bancroft AJ, Else KJ, Grencis RK. Low-level infection with Trichuris muris significantly affects the polarization of the CD4 response. Eur J Immunol. 1994; 24:3113–8. [PubMed: 7805740]
32. Levison SE, et al. Genetic analysis of the Trichuris muris-induced model of colitis reveals QTL overlap and a novel gene cluster for establishing colonic inflammation. BMC Genomics. 2013; 14:127. [PubMed: 23442222]
33. Levison SE, et al. Colonic transcriptional profiling in resistance and susceptibility to trichuriasis: phenotyping a chronic colitis and lessons for iatrogenic helminthosis. Inflamm Bowel Dis. 2010; 16:2065–79. [PubMed: 20687192]
34. Else KJ, Finkelman FD, Maliszewski CR, Grencis RK. Cytokine-mediated regulation of chronic intestinal helminth infection. J Exp Med. 1994; 179:347–51. [PubMed: 8270879]
35. Schopf LR, Hoffmann KF, Cheever AW, Urban JF Jr. Wynn TA. IL-10 is critical for host resistance and survival during gastrointestinal helminth infection. J Immunol. 2002; 168:2383–92. [PubMed: 11859129]
36. D'Elia R, Behnke JM, Bradley JE, Else KJ. Regulatory T cells: a role in the control of helminth-driven intestinal pathology and worm survival. J Immunol. 2009; 182:2340–8. [PubMed: 19201888]
37. Cliffe LJ, et al. Accelerated intestinal epithelial cell turnover: a new mechanism of parasite expulsion. Science. 2005; 308:1463–5. [PubMed: 15933199]
38. Helmby H, Takeda K, Akira S, Grencis RK. Interleukin (IL)-18 promotes the development of chronic gastrointestinal helminth infection by downregulating IL-13. J Exp Med. 2001; 194:355–64. [PubMed: 11489954]
39. Lynch EA, Heijens CA, Horst NF, Center DM, Cruikshank WW. Cutting edge: IL-16/CD4 preferentially induces Th1 cell migration: requirement of CCR5. J Immunol. 2003; 171:4965–8. [PubMed: 14607889]
40. Else KJ, Grencis RK. Antibody-independent effector mechanisms in resistance to the intestinal nematode parasite Trichuris muris. Infect Immun. 1996; 64:2950–4. [PubMed: 8757819]
41. Bos NA, Jiang HQ, Cebra JJ. T cell control of the gut IgA response against commensal bacteria. Gut. 2001; 48:762–4. [PubMed: 11358892]
42. Broadhurst MJ, et al. IL-22+ CD4+ T cells are associated with therapeutic trichuris trichiura infection in an ulcerative colitis patient. Sci Transl Med. 2010; 2:60ra88.
43. Koyama K, Ito Y. Mucosal mast cell responses are not required for protection against infection with the murine nematode parasite Trichuris muris. Parasite Immunol. 2000; 22:13–20. [PubMed: 10607287]
44. Datta R, et al. Identification of novel genes in intestinal tissue that are regulated after infection with an intestinal nematode parasite. Infect Immun. 2005; 73:4025–33. [PubMed: 15972490]
45. Bowcutt R, et al. Arginase-1-expressing macrophages are dispensable for resistance to infection with the gastrointestinal helminth Trichuris muris. Parasite Immunol. 2011; 33:411–20. [PubMed: 21585399]
46. Varin A, Gordon S. Alternative activation of macrophages: immune function and cellular biology. Immunobiology. 2009; 214:630–41. [PubMed: 19264378]
47. Jia TW, Melville S, Utzinger J, King CH, Zhou XN. Soil-transmitted helminth reinfection after drug treatment: a systematic review and meta-analysis. PLoS Negl Trop Dis. 2012; 6:e1621. [PubMed: 22590656]
48. Keiser J, Utzinger J. Efficacy of current drugs against soil-transmitted helminth infections: systematic review and meta-analysis. JAMA. 2008; 299:1937–48. [PubMed: 18430913]

49. Scragg JN, Proctor EM. Further experience with mebendazole in the treatment of symptomatic trichuriasis in children. Am J Trop Med Hyg. 1978; 27:255–7. [PubMed: 646017]
50. Tilney LG, Connelly PS, Guild GM, Vranich KA, Artis D. Adaptation of a nematode parasite to living within the mammalian epithelium. J Exp Zool A Comp Exp Biol. 2005; 303:927–45. [PubMed: 16217807]
51. Knox C, et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res. 2011; 39:D1035–41. [PubMed: 21059682]
52. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012; 22:549–56. [PubMed: 22156294]
53. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–9. [PubMed: 18349386]
54. Myers EW, et al. A whole-genome assembly of Drosophila. Science. 2000; 287:2196–204. [PubMed: 10731133]
55. Hunt M, et al. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 2013; 14:R47. [PubMed: 23710727]
56. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011; 27:578–9. [PubMed: 21149342]
57. Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol. 2010; 11:R41. [PubMed: 20388197]
58. Nadalin F, Vezzi F, Policriti A. GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformatics. 2012; 13(Suppl 14):S8. [PubMed: 23095524]
59. Stanke M, Schoffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 2006; 7:62. [PubMed: 16469098]
60. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011; 12:491. [PubMed: 22192575]
61. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 2008; 18:1979–90. [PubMed: 18757608]
62. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004; 5:59. [PubMed: 15144565]
63. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research. 2003; 13:2178–2189. [PubMed: 12952885]
64. O'Brien KP, Remm M, Sonnhammer EL. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 2005; 33:D476–80. [PubMed: 15608241]
65. Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res. 2011; 39:D289–94. [PubMed: 21113020]
66. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform. 2008; 9:286–98. [PubMed: 18372315]
67. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000; 17:540–552. [PubMed: 10742046]
68. Stamatakis A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22:2688–2690. [PubMed: 16928733]
69. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–2. [PubMed: 20110278]
70. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–9. [PubMed: 19505943]
71. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–11. [PubMed: 19289445]
72. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods. 2013; 10:71–3. [PubMed: 23160280]
73. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–40. [PubMed: 19910308]

74. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11:R106. [PubMed: 20979621]

75. Alexa A, Rahnenfuhrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 2006; 22:1600–7. [PubMed: 16606683]

76. Lynn DJ, et al. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. Mol Syst Biol. 2008; 4:218. [PubMed: 18766178]

77. Quevillon E, et al. InterProScan: protein domains identifier. Nucleic Acids Res. 2005; 33:W116–20. [PubMed: 15980438]



Europe PMC Funders Author Manuscripts

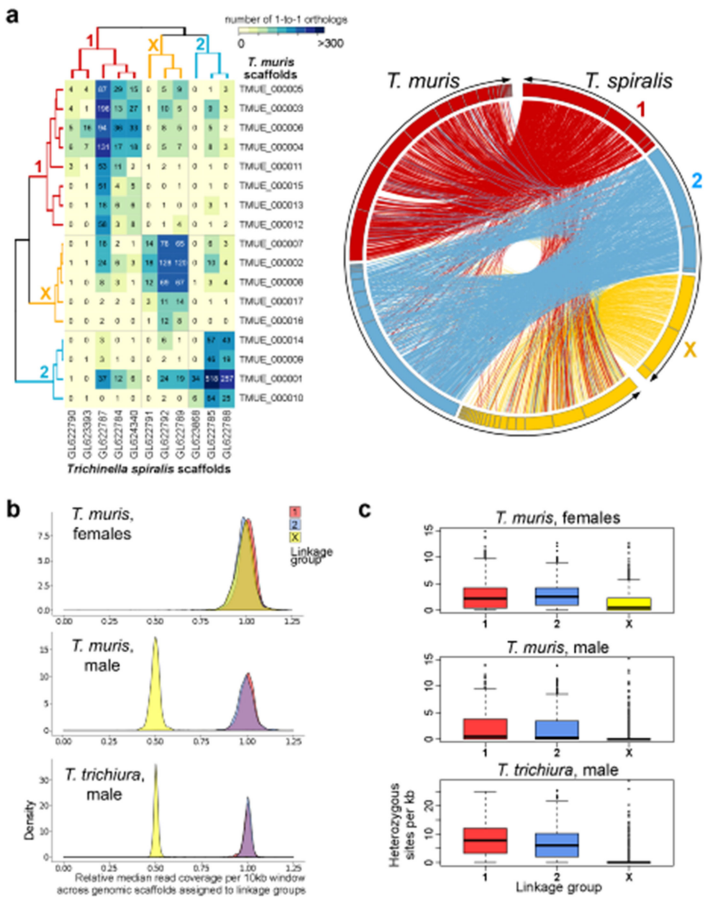


**Figure 1.**

Genome structure and synteny in *Trichinella spiralis* and *Trichuris* spp. (**a**) Mapping one-to-one gene orthologs (as determined by Inparanoid) between genome scaffolds of *Trichinella spiralis* and *T. muris* (genome assembly v4) followed by clustering of the resulting ortholog pattern identifies three large linkage groups in each genome (left; see Supplementary Methods). The table lists the number of one-to-one gene orthologs for individual comparisons between the 11 longest scaffolds of *T. spiralis* and the 17 longest scaffolds of *T. muris*, which represent 85.2% and 89.0% of the respective genomes. The relation of one-to-one gene orthologs illustrates high-level and cross-genus synteny between *T. muris* and *T. spiralis* (right). The linkage group shown in yellow is putatively identified as the sex-specific X chromosome. (**b**) The median relative coverage of high-throughput sequencing reads derived from a pool of 11 female *T. muris* parasites or single male parasites (*T. muris* and *T. trichiura*) and was calculated per 10kb window across all genome scaffolds that were assigned to one of the three linkage groups. In females, mapped sequence read coverage is even across all three linkage groups whereas in males read coverage exhibits a bimodal distribution. In particular, the linkage group to which scaffolds belong separates well with either of the two peaks of relative read coverage. Scaffolds of linkage group X are associated with half the median read coverage found for scaffolds of linkage groups one and two. (**c**)




Europe PMC Funders Author Manuscripts

Levels of heterozygosity correlate strongly with affiliation to one of the three linkage groups. Both the 0.5-fold relative read coverage and the very low apparent heterozygosity of linkage group X are consistent with the corresponding scaffolds representing the sex-specific X chromosome which is expected to occur in a single copy in the diploid genome of a male *Trichuris* parasite.



Europe PMC Funders Author Manuscripts

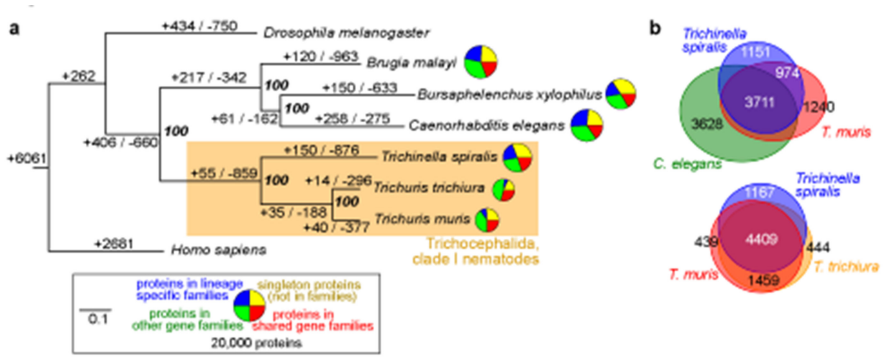


**Figure 2.**

Comparative genomics of *Trichuris*. (**a**) Phylogenetic analysis of genome content. The tree shown is a maximum-likelihood phylogeny based on a concatenated alignment of single-copy orthologs. Values on edges represent the inferred numbers of births (+) and deaths (−) of gene families along that edge. Pie charts represent the gene family composition of each genome – the area of the circle is proportional to the predicted proteome size, and wedges represent the numbers of proteins predicted to be either singletons (i.e. not members of any gene family; yellow), members of families common to the eight genomes (red), members of gene families present only in a single genome (blue), and members of all other gene families (green). (**b**) Euler diagrams of shared presence-and-absence of gene families between clade I nematodes and the model nematode *C. elegans*. Note that the approach in (a) cannot distinguish the polarity of changes at the base of the tree, so for example the value of 262 gene family gains on the basal branch will include gene families lost on the branch leading to *H. sapiens*.




Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

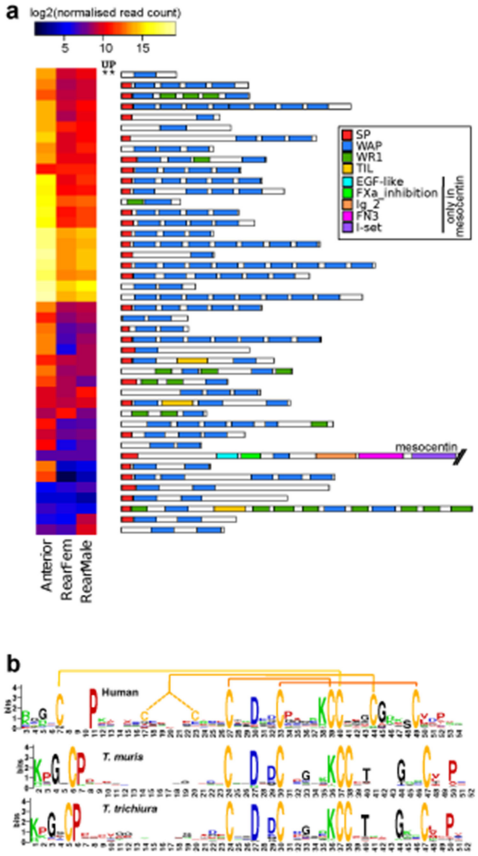


**Figure 3.**

Expression and structural characteristics of WAP domain-containing proteins of *T. muris*. (**a**) Normalised transcript levels of the 44 genes encoding WAP domain-containing proteins in *T. muris* comparing the parasite anterior region with the rear of adult female and male parasites. Abbreviations for protein domain schematics: SP, signal peptide; WAP, Whey Acidic Protein (Interpro IPR008197); WR1, Cysteine-rich repeat (IPR006150); TIL, Trypsin Inhibitor-Like (IPR002919). For a full version of this figure please see Supplementary Fig. 4a. (**b**) The sequence logos illustrate the conserved and distinct sequence characteristics of WAP domains (Interpro IPR008197) found in proteins of *H. sapiens, T. trichiura*, and *T. muris*. The canonical four disulfide bonds formed by eight cysteine residues are highlighted at the top of the sequence logo of the human WAP domains. The sequence logos representing the different species are aligned around the central CxxDxxC motif. For a full version of this figure please see Supplementary Fig. 5.

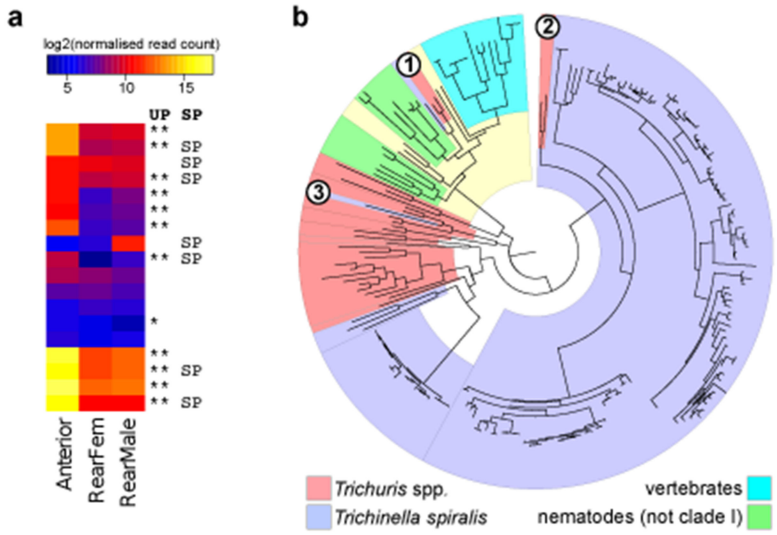


**Figure 4.**

Expression and phylogenetic analysis of DNase II-like proteins of *Trichuris*. (**a**) Normalised transcript levels of the 18 genes encoding DNase II domain-containing proteins in *T. muris* comparing the parasite anterior region with the rear of adult female and male parasites. For a full version of this figure please see Supplementary Fig. 4b. (**b**) A maximum-likelihood phylogeny of DNase II protein domains (IPR004947) illustrates the relationships between DNase II domains of proteins from *Trichuris* spp, *Trichinella spiralis*, other nematodes, insects/other invertebrates, and vertebrates. See Supplementary Fig. 6 for a fully annotated version of this tree. The circled numbers highlight individual sequences of particular interest: (1) TMUE_s0085001500_1-358 (*T. muris*), TTRE_0000372701_1-317 (*T. trichiura*), E5SXW8_TRISP_8-361 (UniProt accession); (2) TMUE_s0015002900_1-191 and TTRE_0000937801_31-255; (3) E5S4S7_TRISP_14-306.

Europe PMC Funders Author Manuscripts





**Table 1**

## Genome and gene statistics for nine nematodes

| | *Trichuris trichiura* [1] | *Trichuris muris* [2] | *Trichuris muris* [3] *(v4)* | *Trichinella spiralis* [4] | *Ascaris suum* [4] | *Brugia malayi* [4] | *Loa loa* [4] | *Meloidogyne hapla* [4] | *Bursaphelenchus xylophilus* [4] | *Caenorhabditis elegans* [4] |
|---|---|---|---|---|---|---|---|---|---|---|
| **Nematode clade** | I | I | I | I | III | III | III | IV | IV | V |
| **Haploid chromosome number** | NA | 3 | 3 | 3 | 12 | 6 | 6 | 16 | 6 | 6 |
| **Genome assembly size [Mb]** | 75.18 | 85.00 | 89.31 [5] | 61.15 | 266.07 | 94.12 | 89.7 | 52.96 | 73.10 | 100.29 |
| **Number scaffolds** | 3,711 | 1,123 | 1,069 | 3,853 | 2,414 | 9,805 | 3,452 | 3,389 | 3,555 | 7 |
| **N50 scaffolds [kb]** | 71.2 | 1,580 | 4,834 | 7,554 | 419.1 | 191.1 | 177.1 | 37.6 | 988.2 | 17,494 |
| **N50 scaffolds [n]** | 263 | 15 | 6 | 3 | 171 | 62 | 126 | 372 | 21 | 3 |
| **Longest scaffold [kb]** | 533.8 | 7,990 | 17,505 | 12,041 | 3,795 | 5,236 | 1,325 | 360.4 | 3,612 | 20,924 |
| **Mean scaffold length [kb]** | 20.3 | 75.7 | 83.5 | 15.9 | 110.2 | 9.6 | 26.0 | 15.6 | 20.6 | 14,327 |
| **Gaps, combined length [kb]** | 14.5 | 227.9 | 4,542 | 4,987 | 7,429 | 7,759 | 3,847 | 0 | 1,476 | 0 |
| **Number genes** | 9,650 | 11,004 | | 16,380 | 15,260 | 14,219 | 14,907 | 14,420 | 18,074 | 20,501 |
| **Mean protein length [aa]** | 435 | 416 | | 318 | 396 | 320 | 334 | 348 | 344 | 404 |
| **Median protein length [aa]** | 329 | 290 | | 192 | 288 | 209 | 213 | 250 | 262 | 329 |
| **Number coding exons** | 55,156 | 83,458 | | 87,853 | 121,416 | 81,154 | 98,974 | 87,957 | 81,552 | 123,608 |
| **Coding exons, combined length [Mb]** | 12.59 | 13.74 | | 15.65 | 18.18 | 13.64 | 14.07 | 15.08 | 18.72 | 24.91 |
| **Mean number coding exons per gene** | 5.7 | 7.6 | | 5.4 | 8.0 | 5.7 | 6.6 | 6.1 | 4.5 | 6.0 |
| **Mean coding exon length [bp]** | 228.3 | 164.6 | | 178.1 | 149.7 | 143.1 | 149.2 | 171.4 | 229.5 | 201.5 |
| **Median coding exon length [bp]** | 148 | 117 | | 129 | 136 | 131 | 136 | 144 | 184 | 146 |
| **Number introns** | 45,506 | 72,454 | | 71,473 | 106,156 | 79,886 | 92,154 | 73,537 | 63,478 | 103,107 |
| **Mean intron length [bp]** | 283.8 | 434.2 | | 197.5 | 1,031 | 626.6 | 669.4 | 153.1 | 153.0 | 201.5 |
| **Median intron length [bp]** | 57 | 58 | | 83 | 726 | 246 | 269 | 53 | 69 | 146 |

All genome statistics are based on scaffolds, filtered for a minimum scaffold length of 1000 bp.

[1] based on *T. trichiura* genome assembly v2 and gene set v2.2, this study.

[2] based on *T. muris* genome assembly v3.0 and gene set v2.3, this study.

[3] based on *T. muris* genome assembly v4.0 which includes super-scaffolds that were joined based on optical map data, this study.

[4] based on genome assembly and gene set release WS236, wormbase.org.

[5] the size shown is directly measured from the assembly. After taking into account the contribution of collapsed repeats, the true size of the haploid female genome is estimated to be 106 Mb (see Supplementary Methods).

## Table 2
## Highly ranked targets with available approved drugs as chemical leads

| Gene ID | Gene annotation | Drugs |
|---|---|---|
| 016011100 | fatty acid synthase | Cerulenin[1], Orlistat[2] |
| 022000400 | Na+,K+ ATPase alpha subunit 1 | Digitoxin[3], Almitrine[4], Bepridil[5], Bretylium[6], Diazoxide[7], Ethacrynic acid[8], Hydroflumethiazide[8], others* |
| 186000500 | serine:threonine protein kinase mTOR | Everolimus[9], Pimecrolimus[10], Sirolimus[9], Temsirolimus[11], Topotecan[11] |
| 016005900 | DNA topoisomerase 1* | Irinotecan[11], Lucanthone[12], Sodium stibogluconate[13] |
| 217000500 | receptor type Tyrosine phosphatase | Alendronate[14], Etidronic acid[14] |
| 010006100 | calmodulin | Aprindine[6], Bepridil[5], Dibucaine[15], Felodipine[16], Fluphenazine[17], Loperamide[18], Phenoxybenzamine[19], others* |
| 069002800 | ribonucleoside diphosphate reductase subunit | Cladribine[11], Gallium nitrate[14] |
| 117003600 | adenosine deaminase | Dipyridamole[20], Nelarabine[11], Theophylline[21], Vidarabine[22] |
| 023008600 | dual specificity mitogen activated protein | Bosutinib[11], Trametinib[11] |
| 029000100 | LDL receptor and EGF-domain containing prot. | Gentamicin[23] |
| 170000300 | integrin alpha pat 2 | Antithymocyte globulin[24] |
| 123000500 | Tyrosine protein kinase Src42A | Dasatinib[11] |
| 013002700 | NADPH cytochrome P450 reductase | Benzphetamine[25], Daunorubicin[11], Ethylmorphine[26], Nitrofurantoin[23], others* |
| 061001100 | DNA polymerase epsilon catalytic subunit A | Cladribine[11] |
| 025003000 | tubulin gamma 1 chain | Vinblastine[11] |
| 092002800 | DNA ligase 1 | Bleomycin[27] |
| 022003300 | amidophosphoribosyltransferase | Fluorouracil[11] |
| 229000900 | V type proton ATPase subunit A | Alendronate[14], others* |
| 012007400 | V type proton ATPase subunit B | Gallium nitrate[14] |
| 031000300 | ADP,ATP carrier protein | Clodronate[14] |
| 304000500 | phenylalanine_4_hydroxylase | Droxidopa[28] |
| 001008300 | proteasome subunit beta type 2 | Bortezomib[11], Carfilzomib[11] |
| 076002300 | Proteasome subunit beta type 5 | Bortezomib[11], Carfilzomib[11] |
| 064003100 | DNA polymerase alpha catalytic subunit | Cladribine[11], others* |
| 225000200 | short:branched chain specific acyl CoA | Valproic Acid[29] |
| 106001600 | histone deacetylase | Aminophylline[30], Lovastatin[31], Vorinostat[11], others* |
| 050001800 | cytochrome c | Minocycline[23] |
| 060007700 | proteasome subunit beta type 1 | Bortezomib[11], Carfilzomib[11] |
| 016000600 | NADH dehydrogenase ubiquinone Fe S protein | Doxorubicin[11] |

Listings of approved drugs obtained from DrugBank[51]. Drug targets have been filtered for predicted essentiality in *C. elegans*, a ChEMBL Ensemble score ≥0.5, transcript expression >100 in adult parasites, and for inferred availability of an approved and possibly suitable drug (excluding e.g. hydroxocobalamin [vitamin B12]; see Supplementary Tables 13, 14). Drugs listed are for the following purposes:

[1] Antifungal

[2] Anti-obesity

[3] Cardiac glycoside (heart failure and arrhythmia)

[4] Respiratory stimulant

[5] Angina treatment

[6] Antiarrhythmic

[7] Vasodilator (hypertension)

[8] Diuretic (hypertension and edema)

[9] Immunosuppressant

[10] Eczema

[11] Cancer

[12] Schistosomicide

[13] Leishmaniasis

[14] Osteoporosis

[15] Local anesthetic

[16] Calcium channel blockers (hypertension, angina, migraine)

[17] Antipsychotic (schizophrenia)

[18] Diarrhea

[19] α-adrenergic-antagonists

[20] Vasodilator (stroke, heart attack)

[21] Asthma

[22] Antiviral

[23] Antibiotic

[24] Organ transplant

[25] Anorectic

[26] Analgesic

[27] Plantar wart (veruca)

[28] neurotransmitter prodrug (hypotension)

[29] Anticonvulsant and mood stabiliser

[30] Bronchodilator

[31] Statin (hypercholesterolemia)

* Other drugs (with similar activities to those listed) are not shown.



Europe PMC Funders Author Manuscripts