Supplementary Information for

Genome of the human hookworm *Necator americanus*

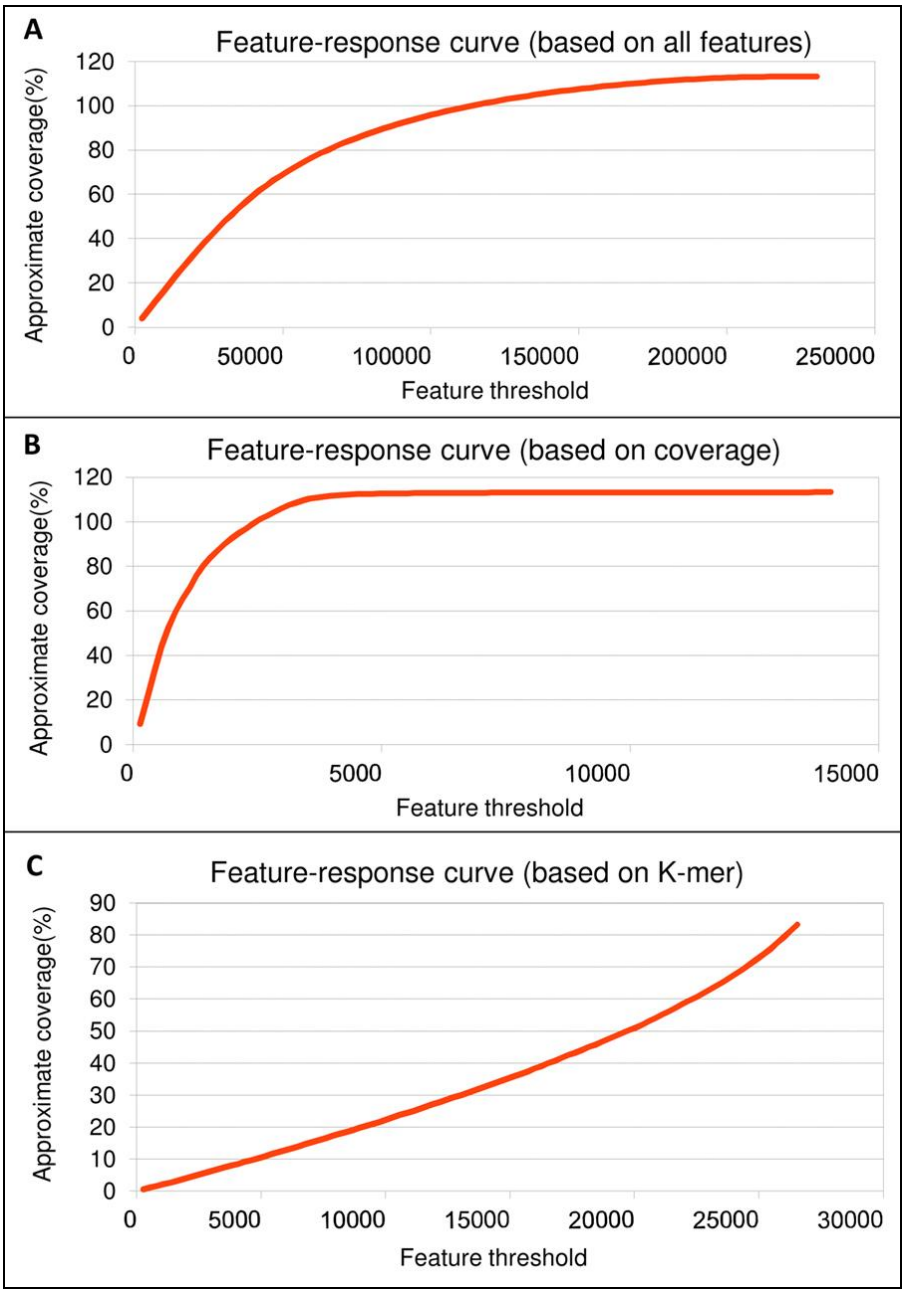
Yat T. Tang[1]*, Xin Gao[1]*, Bruce A. Rosa[1]*, Sahar Abubucker[1], Kymberlie Hallsworth-Pepin[1], John Martin[1], Rahul Tyagi[1], Esley Heizer[1], Xu Zhang[1], Veena Bhonagiri-Palsikar[1], Patrick Minx[1], Wesley C. Warren[1, 2], Qi Wang[1], Bin Zhan[3,4], Peter J. Hotez[3,4], Paul W. Sternberg[5,6], Annette Dougall[7], Soraya Torres Gaze[7], Jason Mulvenna[8], Javier Sotillo[7], Shoba Ranganathan[9,10], Elida M. Rabelo[11], Richard W. Wilson[1, 2], Philip L. Felgner[12], Jeffrey Bethony[13], John M. Hawdon[13], Robin B. Gasser[14], Alex Loukas[7], & Makedonka Mitreva[1, 2, 15#]

*These authors contributed equally to this work
[#]Correspondence should be addressed to mmitreva@genome.wustl.edu

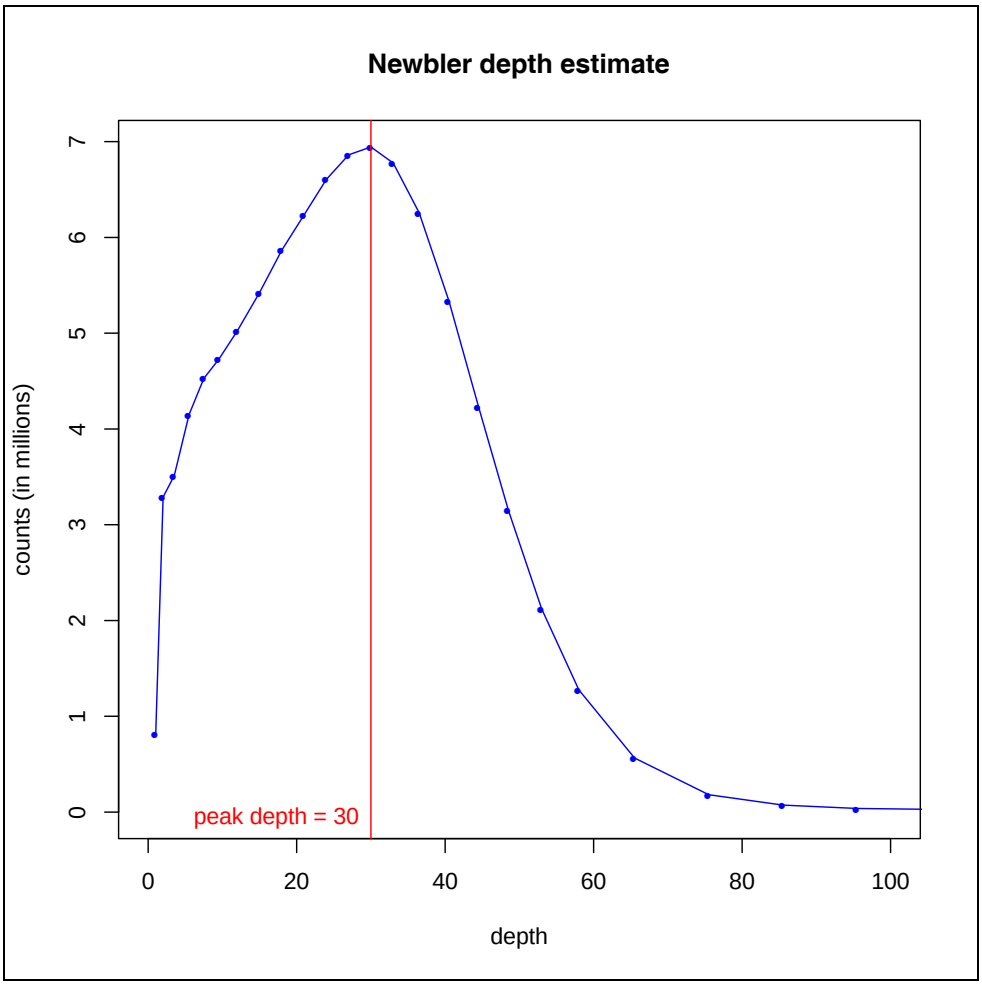
[1] The Genome Institute at Washington University, Washington University School of Medicine, Saint Louis, Missouri, USA.
[2] Department of Genetics, Washington University School of Medicine, Saint Louis, Missouri, USA
[3] Department of Pediatrics, National School of Tropical Medicine, Baylor College of Medicine, Houston, Texas, USA
[4] Sabin Vaccine Institute and Texas Children's Hospital Center for Vaccine Development, Houston, Texas, USA
[5] Division of Biology, California Institute of Technology, Pasadena, California, USA
[6] Howard Hughes Medical Institute, Chevy Chase, Maryland, USA
[7] Centre for Biodiscovery and Molecular Development of Therapeutics, Queensland Tropical Health Alliance, James Cook University, Cairns, QLD, Australia.
[8] Queensland Institute of Medical Research, Brisbane, QLD, Australia
[9] Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, New South Wales, Australia
[10] Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.
[11] Departamento de Parasitologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Minas Gerais, Brazil.
[12] Division of Infectious Diseases, Department of Medicine, University of California Irvine, Irvine, California, USA.
[13] Department of Microbiology, Immunology and Tropical Medicine, The George Washington University, Washington DC, USA
[14] Faculty of Veterinary Science, The University of Melbourne, Parkville, Victoria, Australia.
[15] Division of Infectious Diseases, Department of Internal Medicine, Washington University School of Medicine, Saint Louis, Missouri, USA.

Supplementary Information Index

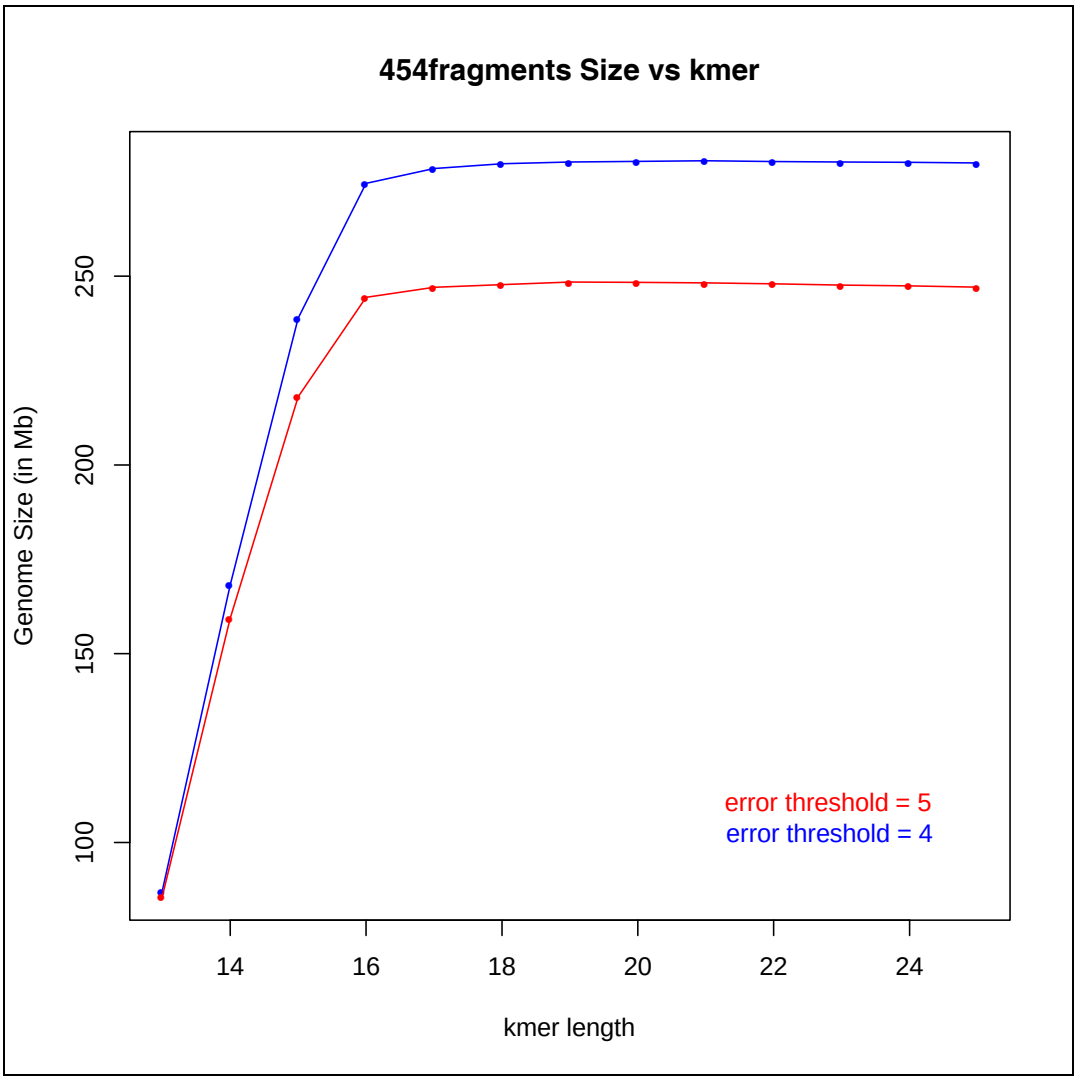
Supplementary Figures 1 to 19



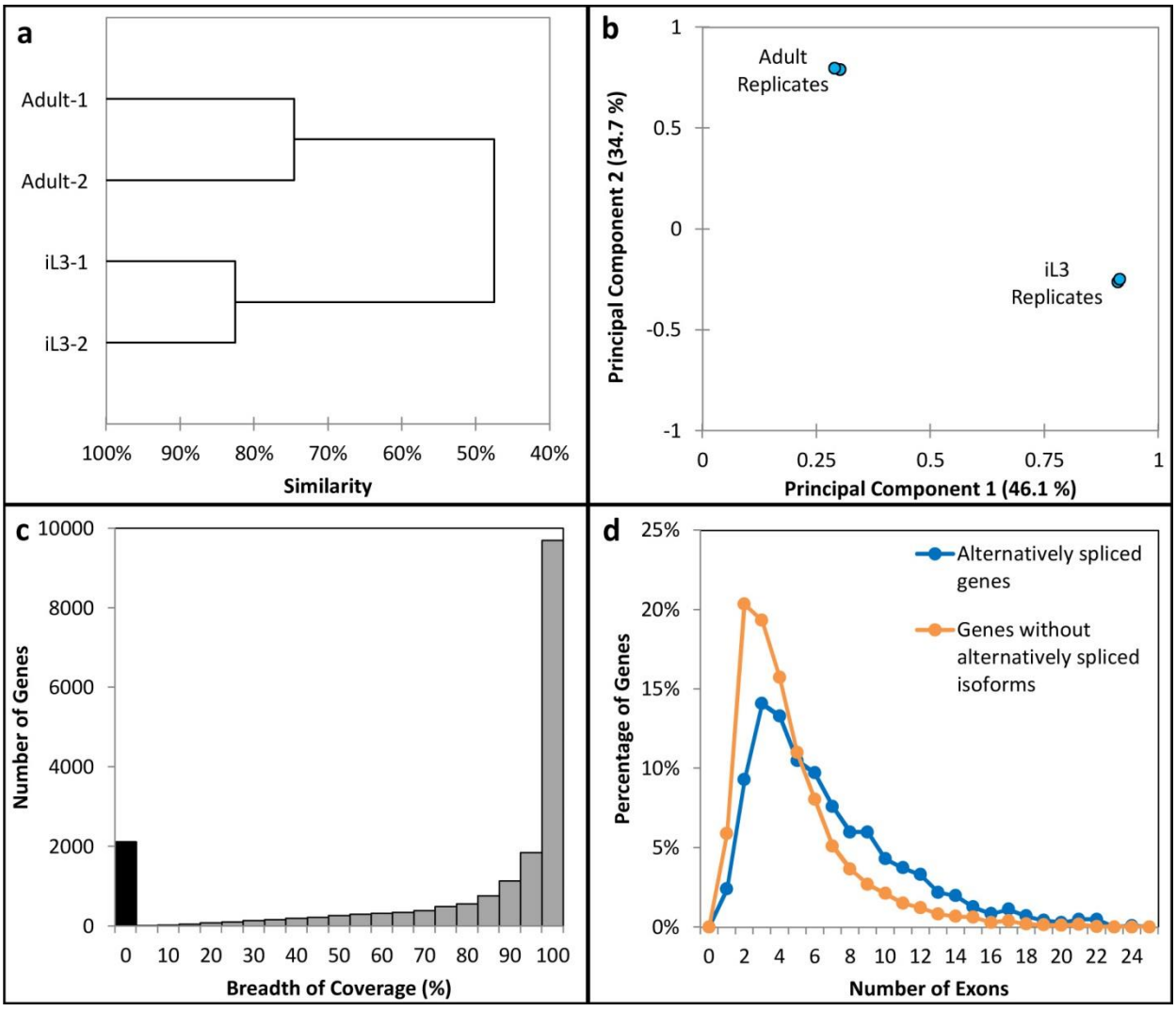
Supplementary Fig. 1: Feature-response curves. Curves were based on (a) all features, (b) coverage and (c) K-mers.



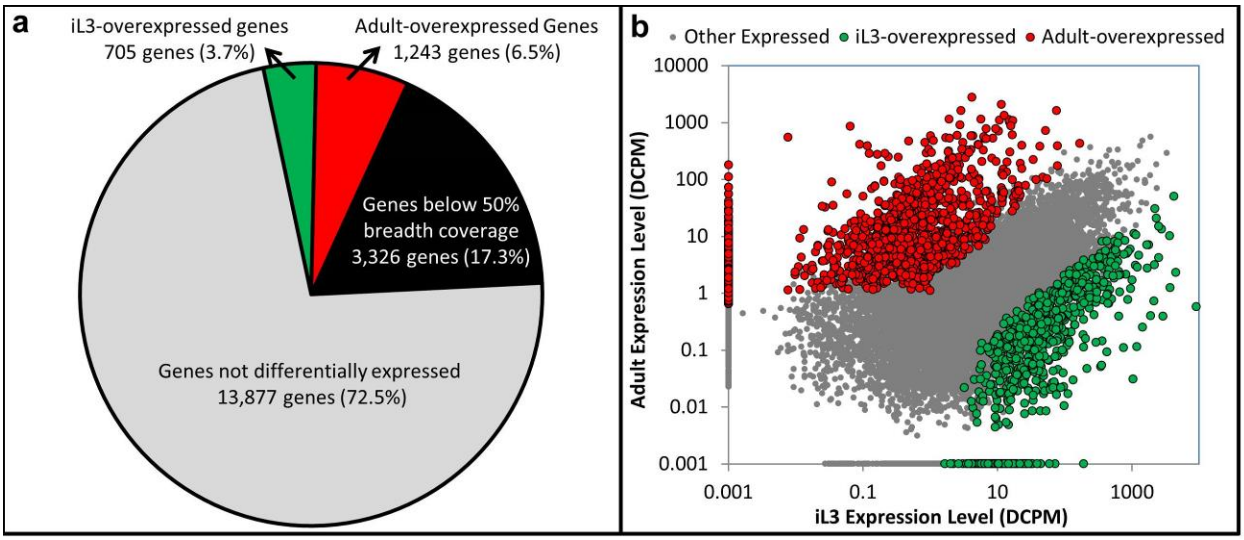
Supplementary Fig. 2: Alignment depth estimation using the depth distribution profile for the Newbler assembly of the *N. americanus* genome.



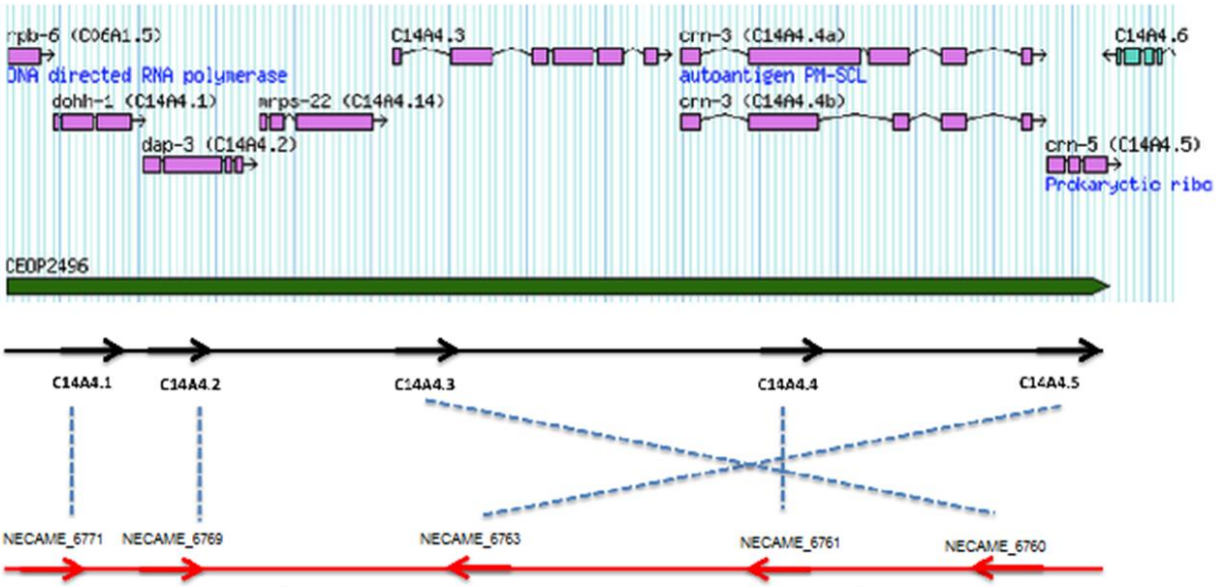
Supplementary Fig. 3: K-mer frequecy-based genome size estimations, using two different error threshold values.



Supplementary Fig. 4. RNA-Seq evidence of the *N.americanus* genes. a. Spearman rank-based hierarchical clustering of RNA-seq samples, based on the expression levels of expressed genes. **b.** Principal component analysis (PCA) plot of of RNA-seq samples, based on the expression levels of expressed genes. **c.** Histogram of the breadth coverage of *N. americanus* genes based on RNA-Seq data (iL3 and Ad). Of the 19,151 genes identified from whole-genome shotgun sequencing, 15,588 genes (81.4%) are confirmed by RNA-Seq data. **d.** Distribution of the percentage of exons in alternatively spliced genes in the *N. americanus* genome. Alternative spliced genes tend to have more exons than genes without alternative spliced forms.

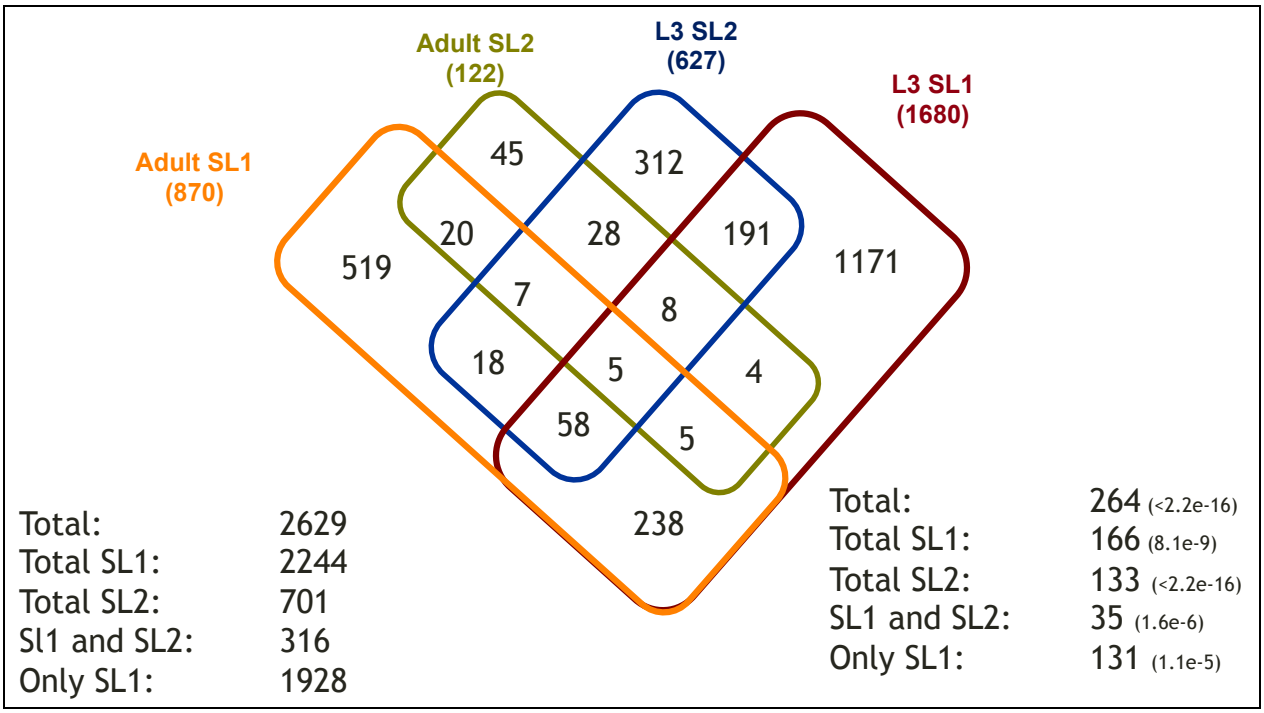


Supplementary Fig. 5. Distribution of *N. americanus* genes based on their expression profiles. a. The number of differentially expressed genes in iL3 (green) and Ad (red), compared to the rest of the *N. americanus* genome. **b.** Expression levels of differentially expressed genes in iL3 (green) and Adult (red) stages.

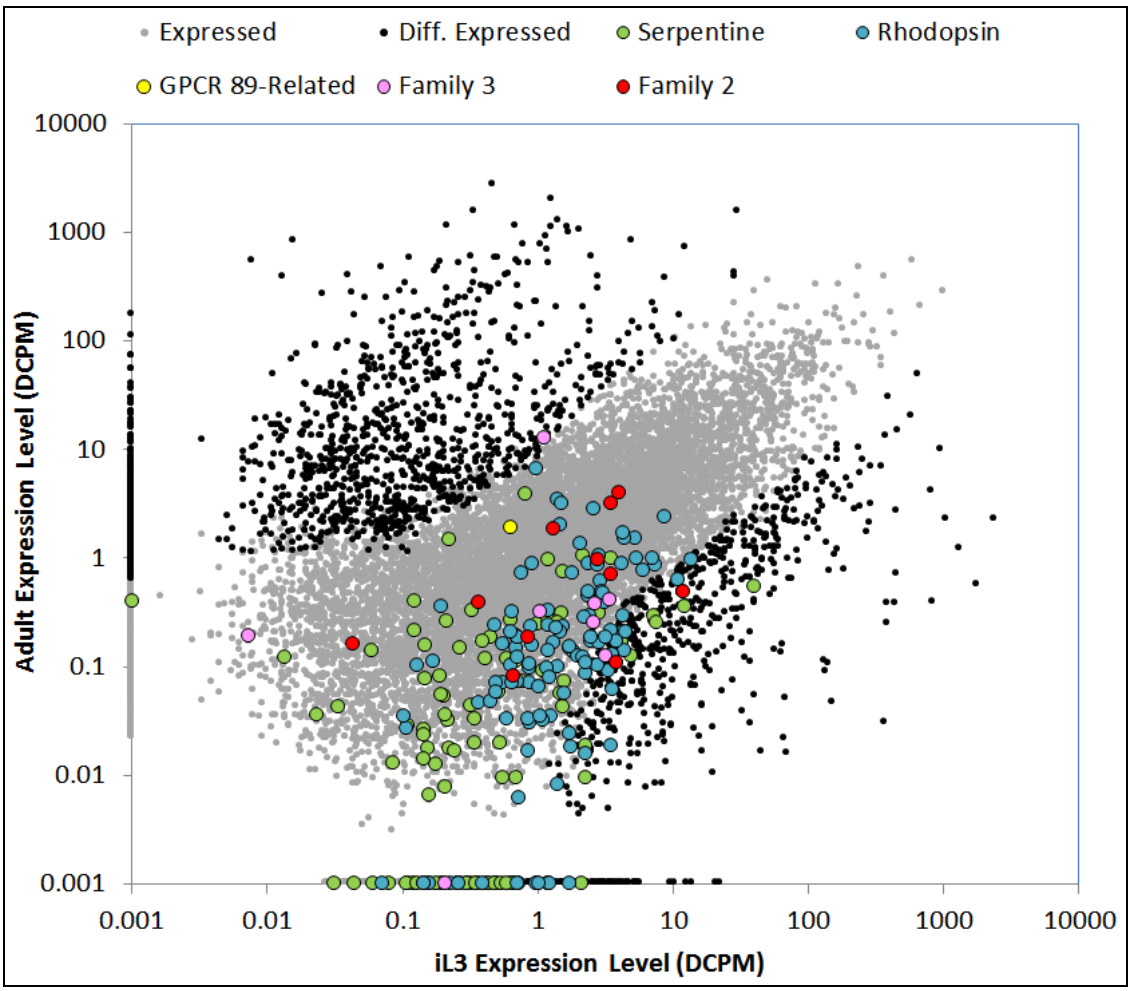


Chromosome	Number of Operons in *C. elegans*	Number of Operons Conserved in *N. Americanus*
I	298	87
II	243	64
III	321	99
IV	225	77
V	192	51
X	77	5

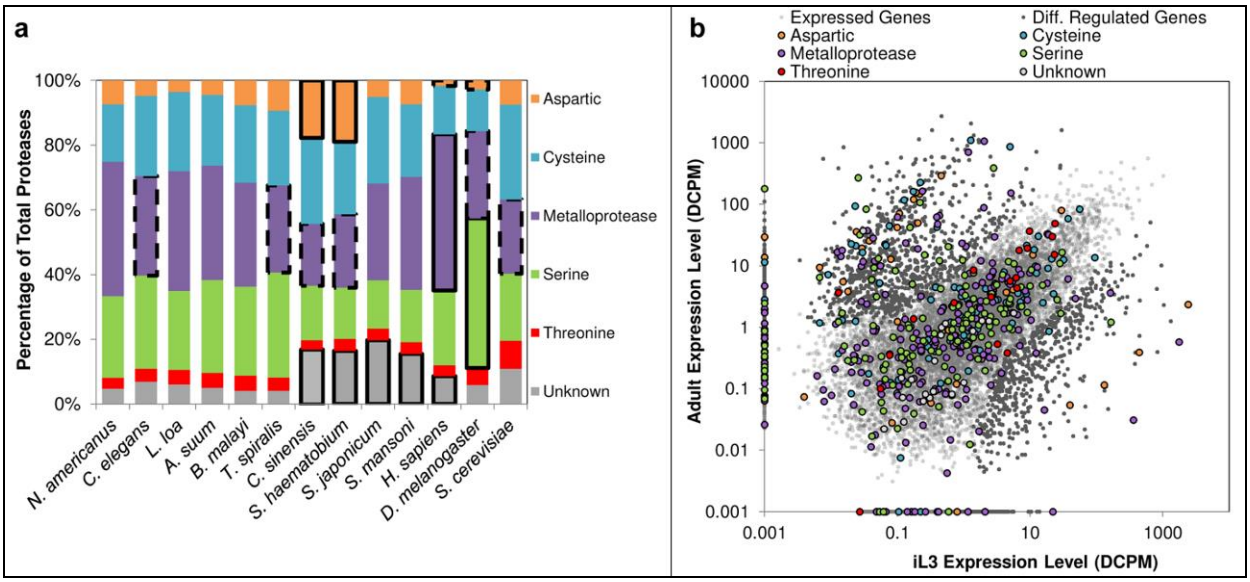
Supplementary Fig.6. Example of an operon with inversion. Distribution of genes within the operons of *C. elegans* (blue) and *N. americanus* (red) are shown for each chromosome. The table shows the distribution of conserved operons over *C. elegans* chromosomes.



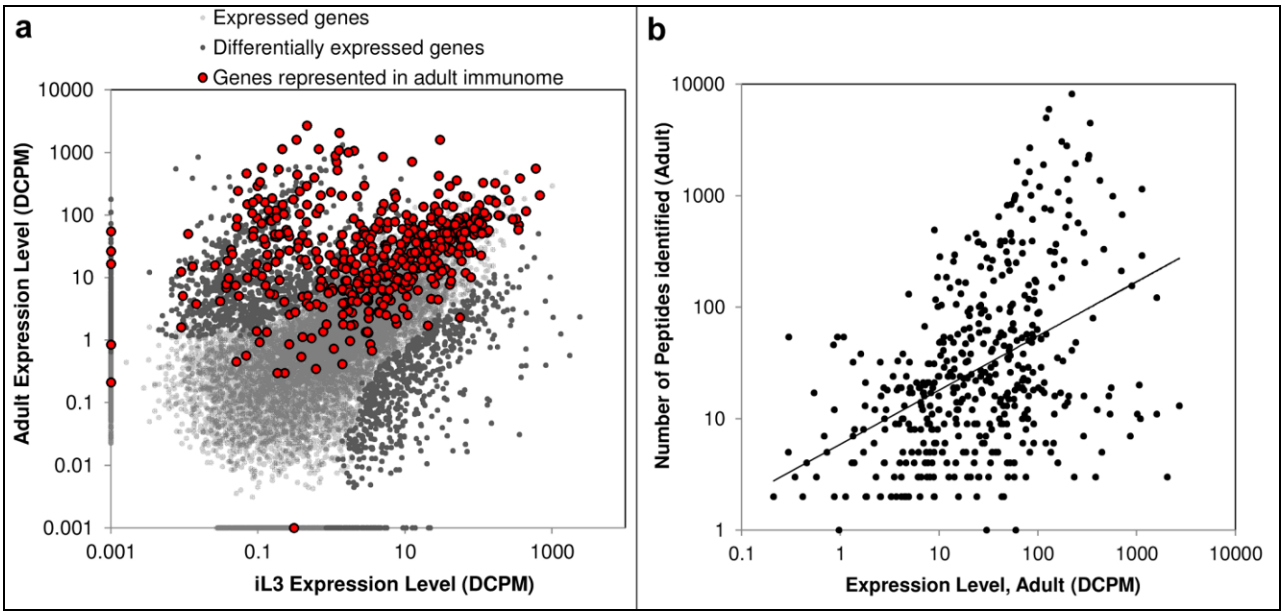
Supplementary Figure 7: The number of genes containing SL1 and SL2 sequences among Adult and iL3 transcripts in *N. americanus*. Total count and significance values compared to orthology-based operon genes are also indicated.



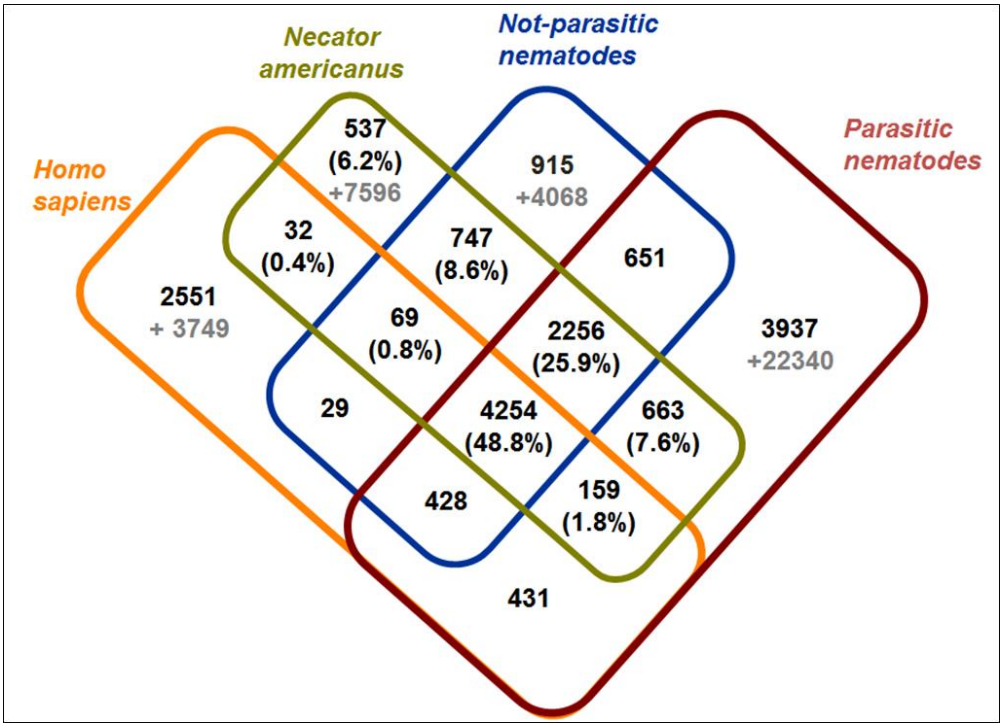
Supplementary Fig. 8. The expression level (DCPM) of the human hookworm *N. americanus* genes containing GPCR Interpro domains (grouped into "Serpentine", "Rhodopsin", "GPCR 89-Related", "Family 3" and "Family 2" domain groups) compared to all other genes.



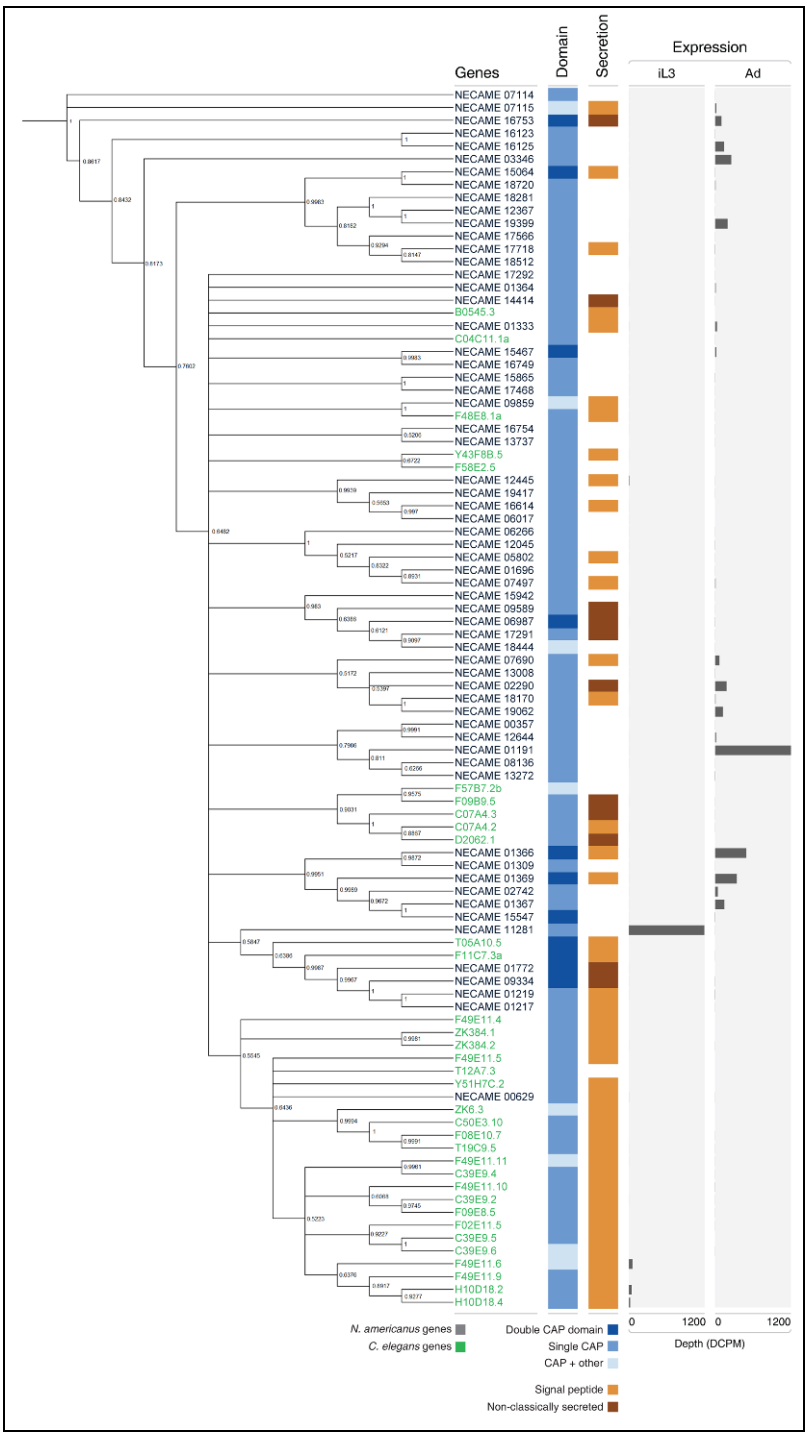
Supplementary Fig. 9. The *N. americanus* degradome. a. Distribution of the major protease classes in the species investigated. Protease families outlined with a solid black line are significantly over-represented compared to *N. americanus* (according to a binomial distribution test), and families outlined with a dashed black line are significantly under-represented compared to *N. americanus*. **b.** The expression level (DCPM) of *N. americanus* proteases in iL3 and adult life cycle stages.



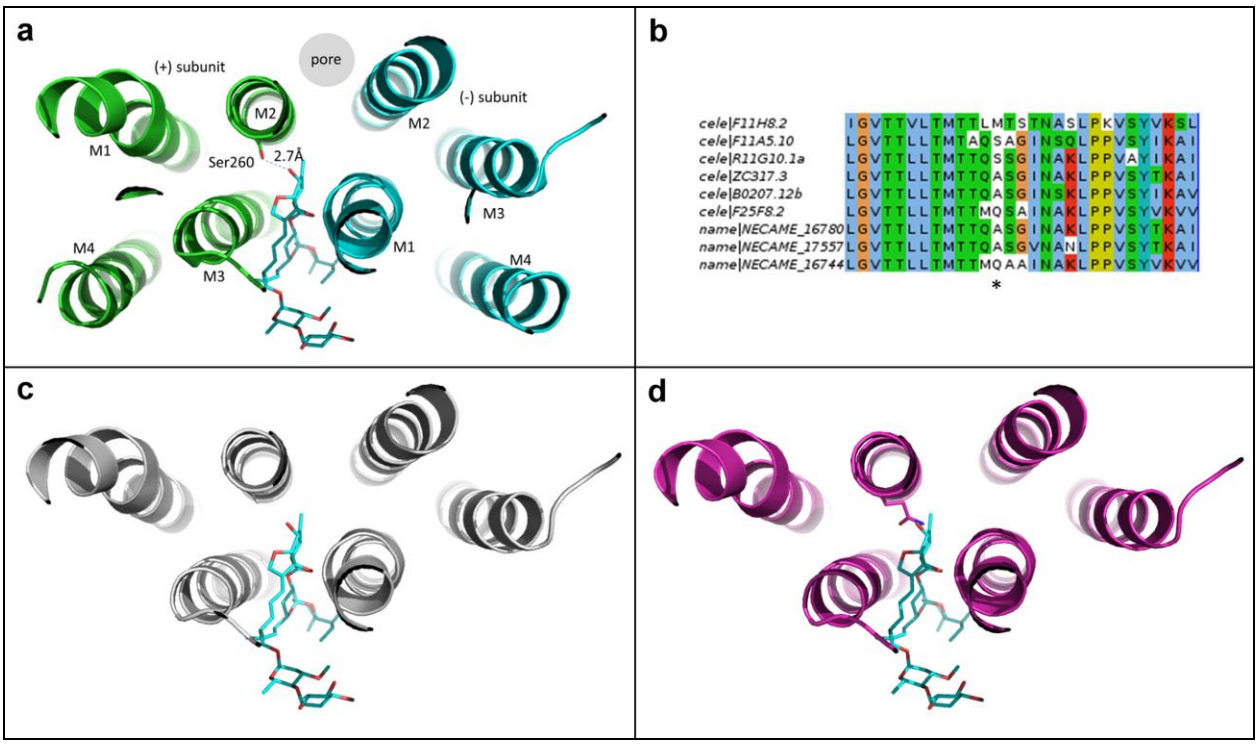
Supplementary Fig. 10. RNA-Seq based expression of genes encoding proteins detected in somatic extract of adult *N. americanus*. a. Adult vs iL3 expression plot showing proteins detected in the Adult-stage specific proteomics dataset. **b.** The number of peptides detected for each protein correlates with gene expression levels for the same life cycle stage on a log scale.



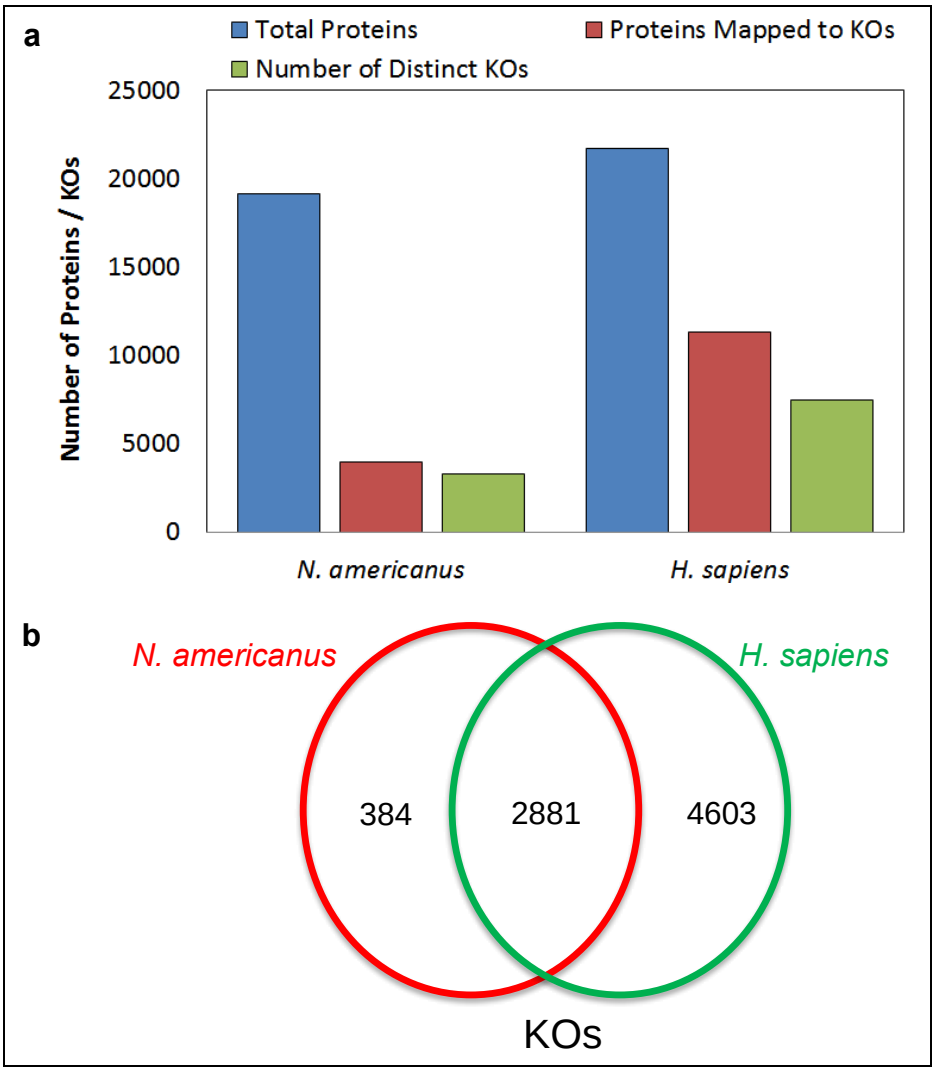
Supplementary Fig. 11. Distribution of orthologous families among *N. americanus,* its host *H. sapiens,* and two nematode phylogenetic databases. The parasitic nematodes database represents the filarial nematodes *B. malayi* and *L. loa*, the large round worm *A. suum* and the zoonotic nematode *T. spiralis*. The non-parasitic nematode is represented by *C. elegans*. Orthologous families were identified using OrthoMCL (inflation factor=1.5). Percentage values indicate the percentage of all orthologous groups in *N. americanus*, and grey numbers (with a "+" symbol) indicate the number of genes which were not included in any orthologous groups.



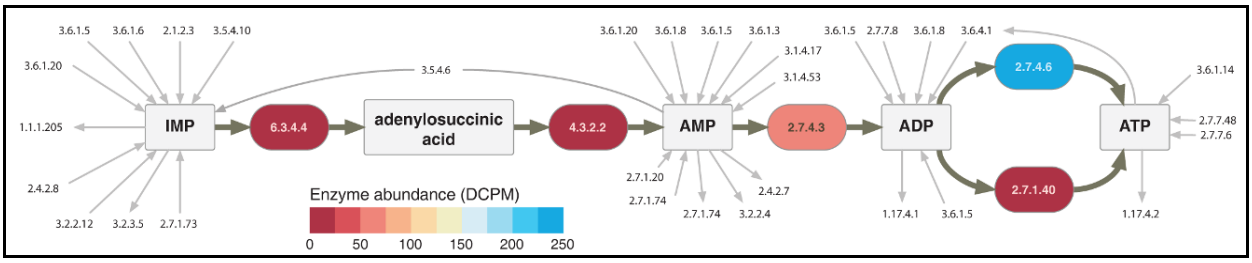
Supplementary Fig. 12: Bayesian-inference clustering of all *C. elegans* and ungapped *N. americanus* SCP/TAPS genes based on their full-length sequence. Data on domain representation, secretion type and stage of expression is included. Numbers within the clustering represent posterior probabilities.



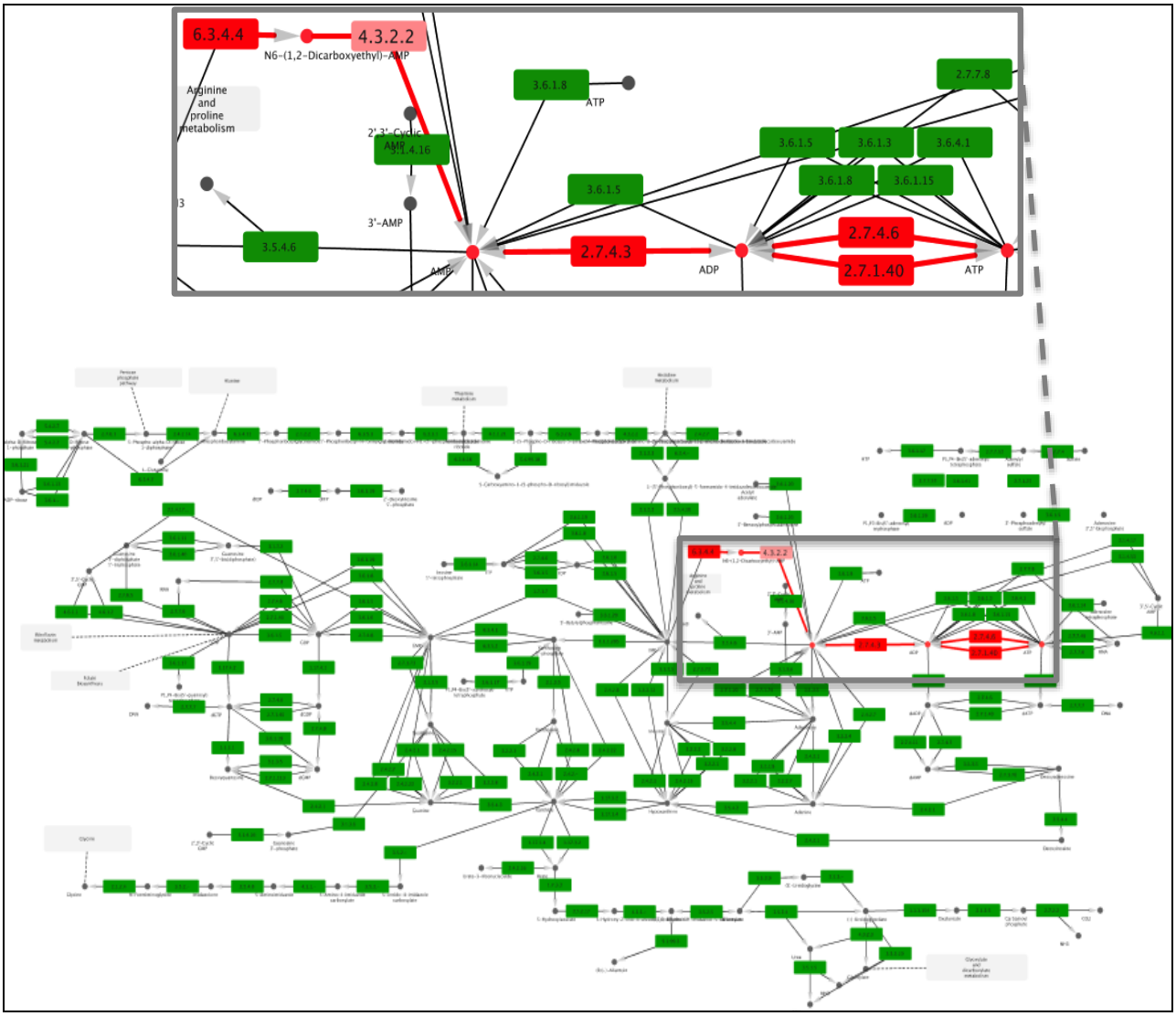
Supplementary Fig. 13. Ivermectin binding at the GluCl channels in *C.elegans* and *N.americanus*. Proteins are rendered in cartoon representation, while ivermectin and Ser260 are shown in stick model. **a.** The hydrogen bond between Ser260 and ivermectin anchors the ligand; **b.** sequence alignments for the *C.elegans* and *N.americanus* orthologs. The position is marked with an asterisk below. **c.** and **d.** Two *N.americanus* ortholog homology models are shown with ivermectin at the same binding site. Either alanine or glutamine is found at the Ser260 position with hydrogen bond disrupted. In both models, only the immediately adjacent helices are shown.



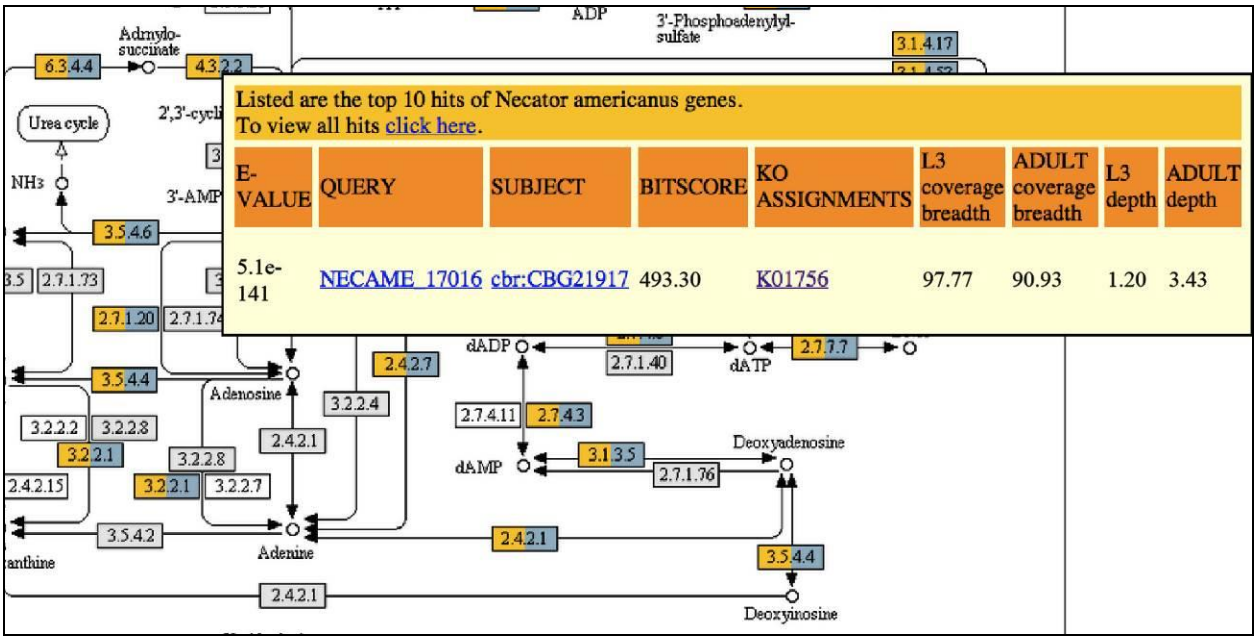
Supplementary Fig. 14. KEGG annotation of the *N. americanus* and *H. sapiens* proteomes. **a.** The standalone version 2 of KEGG Automatic Annotation Server (KAAS) was used to map *N. americanus* and *H. sapiens* proteomes to KEGG Orthology (KO) indices. A blast threshold of 35 bits was used with the bidirectional best hit (BBH) method for KAAS annotation according to the analysis in Tyagi *et al* (in preparation). **b.** Unique KOs shared by the host and parasite proteomes. Only 384 (11.8%) of *N. americanus* KOs are not annotated in the *H. sapiens* proteome. Conversely, majority of the unique KOs (61.5%) mapped to *H. sapiens* proteome were not mapped to *N. americanus* proteome by KAAS.



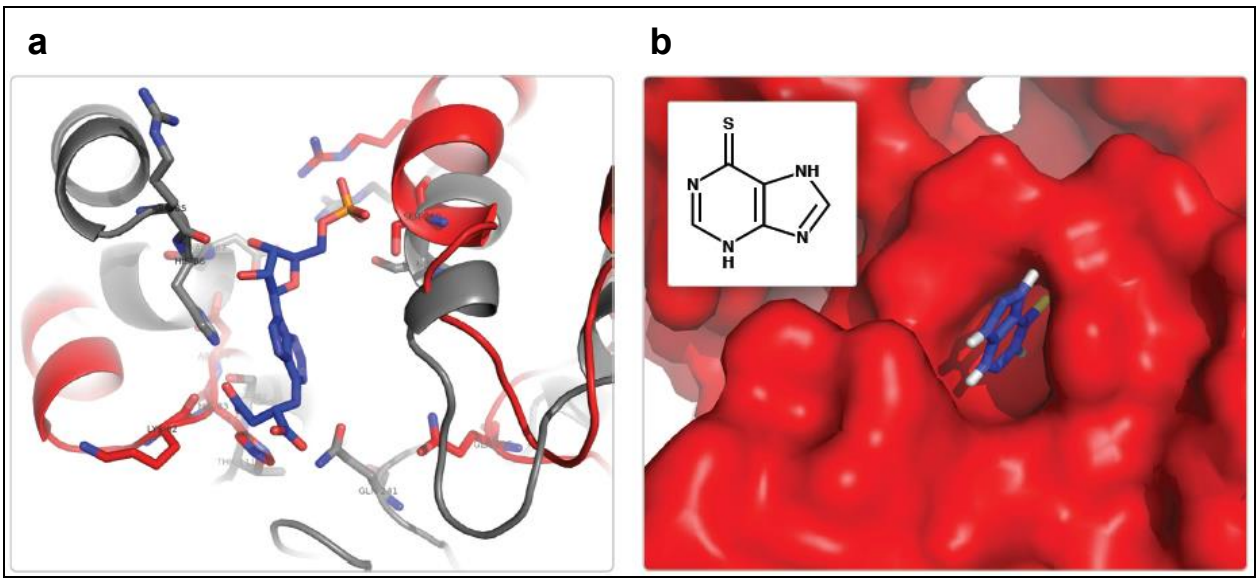
Supplementary Fig. 15. Adenine Ribonucleotide Biosynthesis module (M00049) in the Purine Metabolism pathway (ko00230). *N. americanus* adenylosuccinate lyase (EC: 4.3.2.2) is a chokepoint in the metabolic module for ATP synthesis during the adult parasitic stage of the worm's lifecycle. Colors in the enzyme boxes represent their abundance based on adult stage RNA-Seq data, and while EC: 6.3.4.4. and EC: 4.3.2.2 both show low abundance, EC 4.3.2.2 has slightly lower abundance (3.43 vs 4.46) making it a bottleneck in this module. Targeting a bottleneck enzyme is likely to be more efficient in reducing a metabolic module's output as compared to other enzymes that might be present in excess and hence can tolerate activity reduction better.



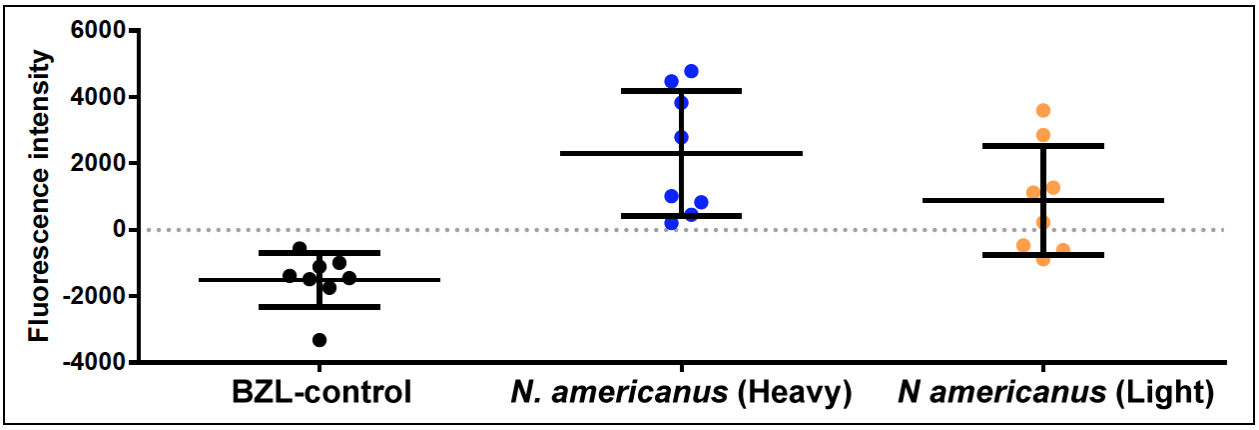
Supplementary Fig. 16. **The purine metabolism pathway (ko00230) with the adenine ribonucleotide biosynthesis module (M00049) shown in red.** Adenylosuccinate lyase (EC 4.3.2.2), the bottleneck and chokepoint of the module, is shown in light red.



Supplementary Fig. 17. A screenshot from the *N. americanus* NemaPath[1] stage-specific pathway viewer. The expression of adenylosuccinate lyase (EC:4.3.2.2) in iL3 and Ad stages in the purine metabolism pathway (ko00230) is shown.



Supplementary Fig. 18. **Homology modeling of *N. americanus* adenylosuccinate lyase and bounding confirmation ofmercaptopurine. a.** Homology model of *N. americanus* adenylosuccinate lyase (EC 4.3.2.2) (red) using the *C. elegans* ortholog (PDB: 1YIS) as the template (98% query coverage, 59% identical, 70% similar), aligned with the human ortholog (gray) (PDB: 2VD6). The substrate, adenylosuccinate (blue sticks), is shown bound with the human ortholog, and the residues involved in binding at the active site (Arg85, His86, Asp87, Thr111, Ser 112, Thr158, Gln241, Ser334, Arg338) are shown for both species. The homology model was built with Modeller[2], and Autodock[3] was used to perform the docking. **b.** Predicted bound conformation of a known inhibitor, mercaptopurine (DB01033) (blue sticks, 2D structure shown in upper left corner), to the homology model of *N. americanus* adenylosuccinate lyase (red).



Supplementary Figure 19: **Immunoreactivity of *NECAME_16652* printed on the microarray.** Fluorescence signal intensity values depicting IgG antibody responses of individual subjects in each cohort to NECAME_16652. The average fluorescence intensity of negative BZL-control sera was subtracted as the background from the raw data and served as the cut-off for immunoreactivity. Only proteins for which a fluorescence signal above the cut off was detected were considered as positive. Error bars represent standard deviation values.

Supplementary Tables

Supplementary Table 1: Amino acid composition (%) of complete proteomes of the 13 species studied.

Amino Acid		Amino Acid Composition in Species (% of All Amino Acids)								
		Nematodes						Host	Outgroups	
		Clade V	Clade V	Clade I	Clade III	Clade III	Clade III			
		N.amer-icanus	*C.ele-gans*	*T.spir-alis*	*B.mal-ayi*	*A. suum*	*L.loa*	*H.sap-iens*	*D.melan-ogaster*	*S.cere-visiae*
F	Phe	4.2	4.8	4.6	3.9	4.2	4.2	3.6	3.4	4.5
S	Ser	7.8	8.1	8.3	8.1	8.3	8.2	8.3	8.6	9.1
T	Thr	5.6	5.9	5.4	5.7	5.5	5.7	5.3	5.8	5.9
N	Asn	4.2	4.9	4.9	4.8	4.3	5.1	3.5	4.8	6.1
K	Lys	5.7	6.4	6.0	6.1	5.4	6.1	5.6	5.5	7.3
E	Glu	6.7	6.5	6.1	6.9	6.7	6.5	7.1	6.5	6.4
Y	Tyr	3.0	3.2	3.1	3.1	3.0	3.2	2.6	2.9	3.4
V	Val	6.7	6.2	6.5	6.1	6.4	6.0	6.0	5.9	5.6
Q	Gln	3.9	4.1	4.3	4.2	3.9	4.2	4.8	5.3	3.9
M	Met	2.6	2.7	2.4	2.5	2.6	2.6	2.1	2.3	2.1
C	Cys	2.2	2.1	2.8	2.1	2.3	2.2	2.3	1.9	1.3
L	Leu	8.9	8.7	9.5	9.2	9.0	9.2	10.0	8.8	9.6
A	Ala	7.1	6.3	6.5	6.3	7.4	6.1	7.1	7.5	5.5
W	Typ	1.1	1.1	1.2	1.1	1.1	1.1	1.2	0.9	1.0
P	Pro	5.1	4.9	4.4	4.6	4.7	4.3	6.4	5.7	4.4
H	His	2.4	2.3	2.6	2.4	2.5	2.5	2.6	2.6	2.2
D	Asp	5.4	5.3	5.0	5.4	5.3	5.4	4.7	5.2	5.8
I	Ile	5.4	6.2	5.6	6.3	5.7	6.5	4.3	4.8	6.6
R	Arg	6.2	5.1	5.9	5.8	6.3	5.8	5.7	5.5	4.5
G	Gly	5.8	5.3	5.0	5.4	5.5	5.1	6.7	6.3	5.0
# Proteins		19151	20517	15840	18348	18542	14893	21860	13917	6692

Supplementary Table 2: *N. americanus* repeat library characterization.

Repeat Class*	Number of Repeat Fragments	Length (bp)
Data from CENSOR		
Interspersed Repeat	2	150
Transposable Element	311	42957
DNA transposon	132	20470
EnSpm	6	461
Ginger2/TDD	1	101
Harbinger	3	156
Helitron	3	241
Kolobok	1	50
Mariner/Tc1	93	17166
Merlin	1	77
MuDR	1	62
Polinton	1	107
Sola	2	125
Zator	2	102
hAT	7	412
piggyBac	1	134
Endogenous Retrovirus	7	436
ERV1	2	162
ERV2	5	274
LTR Retrotransposon	98	11180
BEL	23	2801
Copia	12	658
DIRS	2	489
Gypsy	52	6332
Non-LTR Retrotransposon	73	10749
CR1	12	1100
Crack	2	114
Daphne	1	80
I	1	69
Jockey	2	117
Kiri	1	60
L1	6	388
L2	1	47
L2B	1	83
NeSL	1	47
Penelope	13	1896
Proto1	1	51
Proto2	1	82
R1	1	98
RTE	22	6034
SINE	4	231
SINE3/5S	1	42
Pseudogene	1	58
tRNA	1	58

Data from RepeatMasker	
Low_complexity	518
AT_rich	267
A-rich	191
GC_rich	60
Simple_repeat	129
(TA)n	44
(TCTTG)n	31
(TAAA)n	29
(CAAGA)n	25
SINE/Deu	123
AmnSINE2	123
DNA/TcMar-Mariner	2173
HSMAR1	1987
MARNA	186
tRNA	56
tRNA-Gly-GGY	56
snRNA	125
U4	125

*Specific repeat sequences identified by CENSOR can be downloaded from http://nematode.net/Data/necator_americanus_repeat_library/N_americanus-4.2.2.20110731.RM.1.0.4.lib

Additional Supplementary Tables are provided as separate excel files available for download:

Supplementary Table 3: Differential expression between the iL3 and Adult stages of *N. americanus*.

Supplementary Table 4: Operons in *N. americanus* conserved with *C. elegans* operons.

Supplementary Table 5: Interpro, CAP domain, Gene Ontology (GO) and NCBI Reference (NR) annotations for *N. americanus* genes.

Supplementary Table 6: Stage-specific gene counts and enrichment values for gene ontology (GO) terms.

Supplementary Table 7: KEGG Orthology, transmembrane, signal peptide, non-classical secretion, protease, and kinase classifications for *N. americanus* genes.

Supplementary Table 8: Orthologous Groups for *N. americanus* genes, as classified by OrthoMCL.

Supplementary Table 9: Classification and annotation of genes identified as potential drug targets in *N. americanus.*

Supplementary Table 10: Prioritization of *N. americanus* kinases as drug targets.

Supplementary Table 11: Compounds from DrugBank prioritized as potential kinase drugs.

Supplementary Table 12: Prioritized list of *N. americanus* chokepoints as potential drug targets.

Supplementary Table 13: Complete metabolic modules in *N. americanus* and human and bottleneck genes based on gene expression.

Supplementary Table 14: Compounds from DrugBank prioritized as chokepoints inhibitors.

Supplementary Table 15: Proteins represented and detected (fluorescence intensities) on the protein array.

Supplementary Note

1. Supplementary Methods

Parasite material. The Anhui strain of *N. americanus* was obtained from the laboratory of Dr. Peter Hotez, and was maintained in hamsters[4]. Adult worms were collected from intestines of hamsters infected subcutaneously with *N. americanus* iL3 for 8 weeks[5], and DNA was extracted with the QIAamp DNA Mini Kit according to manufacturer's instruction (Qiagen). For transcriptome sequencing, two key developmental stages from a host-parasite interaction perspective, the infective L3 (iL3) and adult worm, were collected as the representatives of the environmental stages and parasitic stages, respectively. While we made an attempt to collect other stages, obtaining sufficient numbers of individuals from other parasitic stages proved to be problematic because hamsters are not the natural host of *N. americanus*. The transcriptomes that were generated from the two representative stages will likely guide the discovery of targets for drugs, vaccines and diagnostics, as well as novel hookworm-derived protein and peptide biologics for human and veterinary medicine (such as anti-inflammatories and anti-thrombotics). For these reasons, we have chosen to collect these two stages (two biological replicates from each stage) and then apply post-genomic/transcriptomic analyses (e.g. proteomics and immunomics) to these data.

Sequencing, assembly and annotation. Fragment, paired-end whole-genome shotgun libraries (3kb and 8 kb insert sizes) were sequenced using Roche/454 platform. After trimming for linker sequences, the reads were assembled with Newbler[6]. A repeat library was generated using Repeatmodeler and repeats were characterized using CENSOR[7] version 4.2.27 against RepBase (release 17.03[8]). Ribosomal RNA genes were identified using RNAmmer[9] and transfer RNAs

were identified using tRNAscan-SE[10]. Other non-coding RNAs (such as microRNAs) were identified by a sequence homology search against the Rfam database[11]. Repeats and predicted RNAs were then masked using RepeatMasker. Protein-coding genes were predicted using a combination of ab initio programs: Snap[12], Fgenesh, Augustus[13] and the annotation pipeline tool MAKER[14], which aligns mRNA, EST and protein evidence from the same and different species to aid in gene structure determination and modifications. A consensus high confidence gene set from the above prediction algorithms was generated using a logical, hierarchical approach developed at The Genome Institute (described below). Gene product naming was determined by BER (JCVI) and functional categories of deduced proteins were assigned using a suite of protein categorization databases and tools[15-17].

High Confidence Gene Selection. The high confidence gene set was created from the Maker[14] output. First, the following Quality Index (QI) criteria were calculated:

1. Length of the 5' UTR
2. Fraction of splice sites confirmed by an EST alignment
3. Fraction of exons that overlap an EST alignment
4. Fraction of exons that overlap EST or Protein alignments
5. Fraction of splice sites confirmed by a SNAP prediction
6. Fraction of exons that overlap a SNAP prediction
7. Number of exons in the mRNA
8. Length of the 3' UTR
9. Length of the protein sequence produced by the mRNA

Then, the following decision making steps were followed:

a) Genes are screened for overlaps (<10% overlap is allowed).

b) If QI[2] and QI[3] are great than 0, or QI[4] is greater than 0, then the gene is kept.

c) The gene is BLASTed against Swissprot[18] (E<e-6). If there is similarity to a Swissprot entry, then the gene is kept.

d) RPSBLAST is ran against Pfam[19] (E<e-3). If there is similarity to a Pfam entry, then the gene is kept.

e) RPSBLAST is ran against CDD[20] (E<e-3 and coverage >40%). Genes that meet both cut-offs are kept.

f) If no hit is recorded, then a sequence similarity-based search is ran against GenesDB from KEGG[21] , and genes with at least a 55% identity and a bit score of 35 or higher are kept.

Orthologous groups were built from 13 species using OrthoMCL[22] with default parameters. Genes were classified as *N. americanus*-specific if they were not orthologous to any genes in the genomes of the species used as input for this analysis.

Introns and Exons. The size and number of exons and introns in *N. americanus* were determined by parsing exon sizes from gff-format annotations (considering only exon features tagged as "coding_exon") and calculating intron sizes. Due to the draft nature of the genome, genes that were called across gaps in the genomic sequence were omitted from this analysis. Additionally, the exon/intron composition of the *C. elegans* genes (WS230) was analyzed by extracting exons tagged as "coding_transcript" in a similar manner. Due to the maturity of the *C. elegans* genome and geneset, all 20,505 gene loci were used, without considering the status of the underlying gene call. In cases where alternative splicing resulted in multiple transcripts at a

gene locus, the transcript with the largest summed exon size was used to represent that locus. In both species, intron sizes were also averaged on a per-position basis, counting positions considering the 5'-most intron as intron #1, and counting out towards the 3' end.

Significant differences in exon and intron lengths and numbers were tested between species and orthologous and non-orthologous gene groups using two-tailed T-tests with unequal variance. In order to verify normality to justify the use of T-tests, frequency distributions were calculated for the intron sizes, exon sizes, and intron numbers using the values for every gene in both *N. americanus* (ungapped genes only) and *C. elegans*. Then, normal frequency distributions using the average and standard deviations from each of these datasets were calculated, and the Pearson correlation value was calculated to compare the observed and normal distributions. These values were 0.97 for exon sizes (logged), 0.70 for intron sizes (logged), and 0.89 for intron numbers. For *N. americanus* and *C. elegans*, the n values for the T-tests were 11,049 and 20,505 (respectively) when comparing all genes, 4578 and 10,042 (respectively) when comparing orthologous genes and 6471 and 10,463 (respectively) when comparing non-orthologous genes. A p-value significance threshold of (0.05/3)=0.016 was used to correct for multiple testing among the three comparison groups.

Operons. Two separate approaches were used to identify putative operons in *N. americanus*. First, reciprocal best BLAST hits (using WU-BLAST with 30% identity, 35 bits) between *N. americanus* genes and 3,677 *C. elegans* genes (WS230) from operons were used to identify conserved *N. americanus* operons. An operon with at least two *N. americanus* homologs that are adjacent to each other or are separated by one neighbor are considered present. If all of the *C. elegans* genes in the operon have best-hits and all the homologs are contiguous in *N. americanus*,

the operon is considered completely conserved. Stage-specific expression of the conserved operon genes was analyzed using a binomial distribution test for over-representation in the iL3 and adult stages (according to the EdgeR analysis outlined in the "RNA-Seq differential expression analysis" section below). Expression of genes in operon pairs were compared to genes not in operons by sampling 1000 pairs from both sets and comparing the stages that the genes were more expressed in. This was repeated 100 times and the p-value was calculated using a t-test.

The second independent operon identification analysis utilized the known *C. elegans* spliced leader sequences[23] (SL1 and SL2) and the RNA-Seq data that was generated for the adult and iL3 stages of the *N. americanus* life cycle. Reads that satisfied all four of the following criteria for similarity to *C. elegans* SL1 and SL2 sequences were considered to be sourced from a gene trans-spliced with an *N. americanus* spliced leader sequence:

1. A hit was reported by blat, using the options '-tileSize=6 -oneOff=1 -minScore=12'. Matches on either strand were considered hits.
2. The match started, at most, 20 bases from 5' end on the target sequence (RNA-Seq read).
3. The match on the query (the *C. elegans* SL sequence) started at most 3 (for SL1) or 4 (for SL2) bases from 5' end.
4. Match sizes were at least 18 bases (SL1) or 12 bases (SL2); These matches were allowed to spread over more than 1 alignment block.

The matching sequence on the positive strand of reads was used to verify that the sequences obtained had the known characteristics of SL1 and SL2 sequences[23], leading to a higher confidence list of trans-spliced genes. The SL characteristics verified were:

- For SL1: CCCA in the matching sequence followed, after 5 bases, by GAG.

- For SL2: CC[AC][ATG] followed, after 5 to 9 bases, by AAG.

The p-values for enrichment of gene homology based set of operon genes were calculated using Fisher's exact test.

RNA-seq. Roche/454: RNA was extracted[24] from the infective third-stage (iL3) and adult stage of *N. americanus* worms. The integrity and yield of the RNA was verified using the Bioanalyzer 2100 (Agilent Technologies, Cedar Creek, Texas). The same RNA samples were used to generate both Roche/454 and Illumina cDNA libraries. Non-normalized oligo dT libraries for Roche/454 were generated as previously described[24]. The Roche/454 library was sequenced using a Genome Sequencer Titanium FLX (Roche Diagnostics) and the 'sffinfo' program was used to extract information from the SFF files. Adaptor sequences were trimmed from the sequenced reads using the 'seqclean' software and host and bacterial contamination was removed using Newbler's 'gsmapper'. Illumina: For the RNA-Seq library construction, total RNA was treated with Ambion Turbo DNase (Ambion/Applied Biosystems, Austin, TX). 1µg of the DNAse treated total RNA went through poly(A) selection via the MicroPoly(A) Purist Kit according to the manufacturer's recommendations (Ambion/Applied Biosystems, Austin, TX). 1ng of the mRNA isolated was used as the template for cDNA library construction using the Ovation® RNA-Seq version 2 kit according to the manufacturer's recommendations (NuGEN Technologies, Inc., San Carlos, CA). Non-normalized cDNA was used to construct Multiplexed Illumina paired end small fragment libraries according to the manufacturer's recommendations (Illumina Inc, San Diego, CA), with the following exceptions: 1) 500ng of cDNA was sheared using a Covaris S220 DNA Sonicator (Covaris, INC. Woburn, MA) to a size range between 200-400bp. 2) Four PCR reactions were amplified to enrich for proper adaptor ligated fragments and

properly index the libraries. 3) The final size selection of the library was achieved by an AMPure paramagnetic bead (Agencourt, Beckman Coulter Genomics, Beverly, MA) cleanup targeting 300-500bp. The concentration of the library was accurately determined through qPCR according to the manufacturer's protocol (Kapa Biosystems, Inc, Woburn, MA) to produce cluster counts appropriate for the Illumina platform. HiSeq and MiSeq Illumina platforms were used for generation of sequences of 100bp or 150bp from each life-cycle stage.

RNA-seq data were processed using in house scripts. Specifically, BWA was used to detect host contaminated reads that were subsequently removed and DUST[25] was used to identify regions of low compositional complexity that were converted into N's. An in-house script was used to remove reads without at least 60 bases of non-N sequence. RNA-Seq reads from both the iL3 and adult stages were aligned to the predicted gene set using Tophat2.0.0[26] with default parameters (tophat.cbcb.umd.edu). Depth and breadth of coverage measurements for each gene were calculated using RefCov (version 3.0), and expression was quantified using expression values normalized to depth of coverage per mapped bases (DCPM; normalized to 100 million mapped bases). Genes with a breadth of coverage ≥50% across the gene sequence when all reads across all samples were mapped were considered to be "expressed", and were used for further downstream analysis.

Expressed genes were subject to further differential expression analysis using EdgeR[27] (false discovery rate <0.05), in order to identify "iL3-overexpressed" and "adult-overexpressed" genes.

The four RNA-Seq samples were compared for between-sample and inter-replicate variability using standard hierarchical clustering and Principal Component Analysis (PCA) approaches. The expression level (DCPM) for all expressed genes was used as input for both

analyses. For the hierarchical clustering (Supplementary Fig. 2A), the Spearman's rank correlation coefficient was used for similarity calculation, and "Unweighted Pair Group Method with Arithmetic Mean" (UPGMA) clustering was performed (XLSTAT-Pro version 2012.6.02, Addinsoft, Inc., Brooklyn, NY, USA). PCA analysis utilized the Pearson's correlation coefficient, and the first two components (Supplementary Fig. 2B) accounted for 80.8% of the variance between the samples (XLSTAT-Pro version 2012.6.02, Addinsoft, Inc., Brooklyn, NY, USA).

A non-parametric binomial distribution test was used to determine stage-specific enrichment of some groups of genes/proteins (for example, proteins with signal peptides), by testing for enrichment among the genes found to be overexpressed in each stage by EdgeR. Because these tests were ran in pairs (i.e., enrichment was tested among both iL3-overexpressed and adult-overexpressed gene lists), a p-value significance threshold of (0.05/2)=0.025 was used to determine significant enrichment, though all values stated in the manuscript are 0.01 or less. In cases where the p-value was calculated to be equal to zero (according to the BINOM.DIST function in MS Excel 2010), the p-value is given as $p<10^{-15}$, which is the approximate detection limit for the test (dependent on the number of samples in the positive group). For enrichment tests involving more than two comparisons (for example, protease enrichment testing involving different gene groups based on orthlogy), False Discovery Rate (FDR)-corrected binomial distribution probability tests were used[28].

Deduced proteome functional annotation and enrichment. Proteins were searched against KEGG[21] using KAAS[16] (cut-off 35 bits). InterProScan[17] was used to get InterPro[29] domain matches to the proteins using default parameters. Gene Ontology[15] (GO) annotations were

extracted from the InterProScan results. Proteins with signal peptides and transmembrane topology were identified using the Phobius[30] web server, and non-classical secretion was predicted using SecretomeP 1.0[31]. Proteins were also compared to the MEROPS[32] protease unit database using WU-BLAST (with E≤e-10 and only the best hit). Any protein identified as a non-peptidase homolog was ignored. Enrichment of different protease groups among different gene sets (based on similarity to *C. elegans*) was detected based on False Discovery Rate (FDR)-corrected binomial distribution probability tests[28].

The FUNC[33] software package (with default settings) was used to calculate GO enrichment P values. FUNC uses Family-Wise Error Rate (FWER) population correction to population-correct for the number of terms present[33], and only one comparison (iL3→adult) was used as input into FUNC, so no additional population correction was necessary. A 0.01 significance cutoff was used to determine significantly enriched GO categories. QuickGO[34] was used to analyze the hierarchical structure of significant GO categories (Fig. 2a).

Proteomic analysis of somatic worm extract. Whole worms (as described in the "parasite material" subsection of the Methods) were ground under liquid nitrogen before solubilisation in lysis buffer (0.1% (w/v) SDS, 1.0% (v/v) Triton X-100 in 40 mM Tris, pH 7.4). Total protein was precipitated by a 12 hour incubation with nine volumes of methanol at -20^{o}C. Established methods[35] were used to reduce, alkyate and tryptic-digest two 1.5 mg samples of total somatic protein. A 3100 OFFGEL Fractionator™ and OFFGEL Kit pH 3–10 (Agilent Technologies) with a 24-well setup were prepared as per the manufacturer's protocols. The tryptic digests were diluted in water (without the addition of ampholytes) to a final volume of 3.6 ml and 150 µl was loaded into each well. The samples were focused with a maximum current of 50 µA until 50 kVh

were achieved. Peptide fractions were harvested, lyophilised and resuspended in 5% formic acid before LC and mass spectral analysis.

Tryptic fragments from in-gel digests were chromatographically separated on a Dionex Ultimate 3000 HPLC using an Agilent Zorbax 300SB-C18 (3.5 µm, 150 mm x 75 µm) column and a linear gradient of 0-80% solvent B over 60 min. A flow rate of 0.3 µl/min was used for all experiments. The mobile phase consisted of solvent A (0.1% formic acid (aq)) and solvent B (80/20 acetonitrile/0.1% formic acid (aq)). Eluates from the RP-HPLC column were directly introduced into the NanoSpray II ionisation source of a 5600 MS/MS System (AB Sciex), operated in positive ion electrospray mode. All analyses were performed using Information Dependant Acquisition. Analyst 2.0 (Applied Biosystems) was used for data analysis and peak list generation. Briefly, the acquisition protocol consisted of the use of an Enhanced Mass Spectrum scan as the survey scan.

The three most abundant ions detected over the background threshold were subjected to examination using an Enhanced Resolution scan to confirm the charge state of the multiply charged ions. The ions with a charge state of +2, +3 or with unknown charge were then subjected to collision-induced dissociation using a rolling collision energy dependent upon the m/z and the charge state of the ion. Enhanced Product Ion scans were acquired, resulting in full product ion spectra for each of the selected precursors, which were then used in subsequent database searches.

Searches were performed using version 12.10.01.1 of X!Tandem[36] with a 0.4 Da tolerance on the precursor, 0.5 Da tolerance on the product ions, allowing for methionine oxidation and carbamidomethylation as fixed and variable modifications (rxespectively, one missed cleavage, charge states +2, +3 and +4 and trypsin as the enzyme. All experiments were

searched against the *N. americanus* predicted protein dataset (19,151 proteins). Searches were also made against the SwissProt database (updated April 23, 2013) to control for contamination, and incorporated decoy searches[37] against reversed databases and a semi-supervised learning algorithm was used to screen peptide identifications and impose a false discovery cut-off of 1% using the "Self-boosted Percolator" algorithm[38]. Peptide identifications were then grouped into a parsimonious set of identifications and protein probabilities calculated using ProteinProphet[39]. Only proteins identified with at least two peptides and a confidence of $p \leq 0.05$ were considered identified (corresponding to a false discovery rate of 0.0%).

FUNC[33] was used to test for functional enrichment among the genes supported by proteomics from this dataset, using the same settings as outlined in the "Gene Ontology Functional Enrichment" subsection, using all of the genes without proteomics support as a background for comparison.

Transcription Factors and the binding sites. The transcription factors were identified using the KEGG transcription factor database that documents a total of 833 transcription factor records collectively from eukaryotic and prokaryotic organisms (derived from TRANSFAC 7.0[40]). KEGG Orthology (KO) numbers from this database were intersected with the KOs associated with *N. americanus* proteins, to build *N. americanus* database of transcription factors.

Documented matrices of transcription factor binding sites were downloaded from the JASPAR database[41]. The corresponding protein accession numbers were extracted and converted to KOs. Comparison of *N. americanus* transcription factor KOs and those KOs defined a subset of *N. americanus* transcription factors with available binding site information. The binding site matrices of this subset of *N. americanus* transcription factors were used to scan the sequences of

up to 500 bp downstream and upstream of differentially expressed genes using Patser software with default parameters.

SCP/TAPS. SCP/TAPS (CAP) domain-containing sequences were identified by searching each protein for the SCP/TAPS-representative protein domains[42] IPR014044 ("CAP domain") and PF00188 ("CAP")[43] using Interproscan[17] and hmmpfam[44] (respectively) with default parameters. The domain ID, number of copies and locations for each protein sequence were recorded. Phylogenetic relationship trees using the full length of primary sequences derived from the ungapped genes were built using Bayesian inference[45] and Neighbor Joining[46] methods in a similar manner to recent large-scale studies analyzing SCP/TAPS from a range of parasitic helminths[42,43,47]. Leaves of the tree were annotated with domain information, secretion mode and expression data, and then visualized using iTOL[48].

Potential Drug Targets. GPCRs, LGICs, and VGICs were identified based on searches in InterProScan. Genes classified as "7TM GPCR serpentine receptor class" in InterProScan are considered as potential chemoreceptors. Nuclear receptors and neuropeptides were identified using WU-BLASTP (E≤e-10) against the *C. elegans* proteome (WS230). Ivermectin target characterization analysis: Sequence alignments were obtained by MUSCLE[49] for the *C. elegans* and *N. americanus* orthologs within two orthologous groups (NAIF1.5_00184 and NAIF1.5_06724). Homology models for the two *N. americanus* orthologs (*NECAME_16744* and *NECAME_16780*) were built by MODELLER[2] using the *C. elegans* crystal structure as template[50]. For each ortholog, five models were built and the one with the lowest total function score (energy) was chosen as the model shown. Sequence alignments are colored by Clustalx

scheme in JalView[51]; protein structure models are rendered in PyMol (Schrodinger, LLC, The PyMOL Molecular Graphics System, Version 1.3r1. 2010).

Kinome. *N. americanus* genes were screened against the collection of kinase domain models in the Kinomer[52] database v1.0. Custom score thresholds for each kinase group were taken from Kinomer and then adjusted until an hmmpfam search (HMMER 2.3.2)[44] came as close as possible to identifying known *C. elegans* kinases using the Kinomer allPK.hmm profile database. Those same cutoffs were then applied to the *N. americanus* gene set to identify putative kinases as previously descibed[53]. Manual curation was then performed by screening the putative kinases against Pfam[54] using hmmpfam (as a part of an InterProScan run [InterProScan software v4.8, Pfam db release 24.0]). Putatively identified kinases found not to have a kinase or kinase-related Pfam domain hit were removed from the final set of kinase identifications[55].

Kinase prioritization. Kinase prioritization was done as previously described[56]. In summary, proteins were grouped into orthologous groups (see "Identification of orthologs through markov clustering" section). To identify the most appropriate groups of kinases for which there are targeting compounds, the kinase groups (based on the OrthoMCL clustering) were evaluated for RNAi phenotype and tissue expression data in *C. elegans*. RNAi phenotypes for *C. elegans* were classified according to Kumar et al.[57] and the complete list of RNAi phenotypes (sorted by bin) were used. Tissue expression for *C. elegans* was obtained from Wormmart[58]. GO associations of all helminth proteins were inferred using InterProScan[59]. These GO associations, along with transcript expression, tissue localization[58], and *C. elegans* RNA-Seq data[60] were used to evaluate potential groups for compound testing[56,61].

Chokepoint identification, prioritization, module completion and bottlenecks. Chokepoints of KEGG metabolic pathways were identified based on the following criteria. A "chokepoint reaction" was defined as a reaction that either consumes a unique substrate or produces a unique product. The reaction database from KEGG v58[62] was used. Each reaction equation is listed as a separate reaction with a unique identifier under the ENTRY field. The KEGG reaction database also contains a file that lists the reactions within the reaction database as reversible or irreversible (reaction_mapformula.lst – downloaded 6/21/2011). The entire reaction was extracted from the KEGG reaction database by parsing the EQUATION field, and the reaction_mapformula.lst file was used to obtain the directionality of the reaction such that the reactions could be written with reactants on the left side and products on the right side. If the reaction was reversible, this was also noted in the file because products and reactants would be ambiguous. The reactions were placed into a [compound x reaction number] matrix by parsing an intermediate file that contained the directionality and all of the products and reactants for the reaction within the matrix; -1 indicated the compound was consumed (i.e. the compound was listed on the left side of the equation), +1 indicated the compound was produced (i.e. the compound was listed on the right side of the equation), 2 indicated the reaction was reversible, and a zero indicated the compound did not take part in the reaction. To find the chokepoints, the matrix was parsed for compounds that were only produced or consumed in a single reaction. If a compound was produced or consumed in a single reaction, only a single 1 or -1 would be present across the entire compound row within the matrix. In some cases, a compound was uniquely produced or uniquely consumed, but was part of a reversible reaction (i.e. two 2's would be present within a row). If this reaction was the only reaction in which the compound participated,

this was also called a chokepoint. The chokepoint compounds were related to EC numbers (Enzyme Commission number) using the ENZYME field in the reaction database.

KEGG Orthology assignments were performed based on a standalone version of the KEGG Automatic Annotation Server (KAAS)[16] with a BLAST bit-score threshold of 35 in the bi-directional best hit (BBH) method. Module abundances were calculated, and module bottlenecks defined as reaction steps in the cascade that are both essential for the module completion and have low enzyme abundance that primarily constrains the overall module abundance (Tyagi et al, unpublished). Enzyme abundances were calculated using the iL3 and adult stage RNA-Seq expression levels (DCPM).

EST sequence sets were downloaded from Genbank on 7/16/2010. *C. elegans* EST sequences were downloaded from GenBank on 6/15/2012. The tissue expression data from *C. elegans* was obtained from WormMart (WS210) on 6/15/2012.

Homology models were aligned with their reference sequence using T-COFFEE[63], constructed using MODELLER[64] with default parameters using PDB structures with the highest sequence similarity, and docking was performed using AutoDock4.2[3] using default parameters (Supplementary Fig. 14).

Prioritization of chokepoint reactions and targets was performed by assigning a point for meeting each of the following criteria, then ranked based on number of points: transcript based expression found in a infective/parasitic stages; expression in specific tissues (pharynx, intestine, neurons, muscle, or hypodermis) in *C. elegans*; less than 30% sequence identity to *H. sapiens* over half the length of the sequence; chokepoint enzyme functioning in two or more pathways; chokepoint enzyme involvement in nucleic acid metabolism; and chokepoint being a hydrolase based on their enrichment (classification as EC 3) as targets in the drug databases.

Chemogenomic screening for compound prioritization. Proteins associated with ECs (using KEGG) were BLAST searched against protein targets in DrugBank as described above to identify targets with known inhibitors or approved drugs. Cheminformatic properties were obtained by running SMILES strings (SMILES are strings of ASCII characters that describe a compound unambiguously) extracted from DrugBank[65] through the Cytoscape[66] plugin, ChemViz. To evaluate current drugs and inhibitors that target homologous kinases, compounds from a publicly available database (DrugBank[65]) were prioritized based on physico-chemical properties favorable for pharmaceutical compounds[55]. One point was given to each compound for meeting each of the following criteria: (1) molecular weight $\leq$ 500; (2) 0 < number of rotatable bonds $\leq$ 10; (3) hydrogen-bond donors $\leq$ 5; (4) hydrogen-bond acceptors $\leq$ 10; (5) logP $\leq$ 5[55]; (6) half-life $\geq$60 minutes; (7) toxicity information available. The maximum attainable score for top prioritized compounds is 7. Drug-like compounds were also eliminated if placed in the dietary supplement, micronutrient, or vitamin categories by DrugBank, as various vitamins and amino acids were not desired. Proteins were searched against sequences from DrugBank, and then parsed for sequences that had 50% or greater identity over 80% of sequence length. Only these targets were considered in the prioritized list.

Protein microarray. Study Sample: In 2005, 1494 individuals between the ages 4 and 66 years (inclusive) were enrolled into a cross-sectional study in an N. americanus-endemic area of Northeastern Minas Gerais state in Brazil, using protocols approved by the George Washington University IRB (117040 and 060605), the Ethics Committee of Instituto René Rachou, and the National Ethics Committee of Brazil (CONEP) (Protocol numbers 04/2008 and 12/2006).

Informed consent was obtained from all subjects. Individuals were determined to be positive for N. americanus by light microscopy using an ether-sedimentation technique. If positive, the intensity of their hookworm infection was quantified using "eggs per gram of feces" (epg), as measured using the Kato-Katz fecal thick smear technique (with 2 slides from 3 consecutive days of fecal samples). Venous blood (15 mL) was collected into siliconized tubes and the sera were separated, aliquoted, and stored at -80°C in temperature-monitored freezers. For the current pilot protein array analyses, the study sample included individuals who were positive for N. americanus infection based on the ether-sedimentation technique. The egg positive individuals were then stratified by 15-year age intervals (starting at age 4) and randomly sampled within each age strata using proportional allocation (i.e., the sample size in a stratum was proportional to the number of units in the stratum).

Protein Microarray construction and probing. *N. americanus* open reading frames (ORFs) were selected for expression and printing from the predicted proteins from the genome sequence if they contained a classical signal peptide (as determined using the SignalP v4.1 server) and if there was evidence for the ORF's transcription in iL3 and/or adult worms from supporting RNA-Seq data. Of the 1,275 predicted proteins that fit these criteria, 623 corresponding cDNAs were successfully amplified by PCR for *in vitro* transcription and translation. Primer design and PCR amplification from iL3 and adult worm cDNA libraries were performed as previously described[67]. Amplicons were cloned into the pXi T7 vector by homologous recombination as previously described[68]. Resultant plasmids were purified and the inserts were verified by PCR and sequencing. Plasmids with correct inserts were expressed in an *in vitro* cell-free system based on *Escherichia coli* ribosomes (Roche RTS 100), and the protein extracts were contact-

printed without purification onto nitrocellulose glass FAST® slides. As controls, adult and iL3 somatic extracts of *N. americanus* were spotted onto the array. Non-hookworm control proteins/RTS reactions were also spotted onto the microarray as described[67]. The printed *in vitro*-expressed proteins were quality checked using antibodies against incorporated N-terminal poly-histidine (His) and C-terminal hemaglutinin (HA) tags.

Probing of protein microarrays with human sera. Protein arrays were blocked in blocking solution (Whatman) for 2 hours at RT prior to probing with human sera (diluted 1:50) at 4^{o}C overnight. Arrays were washed 5 times in TBS-containing 0.05% Tween-20, then isotype and subclass-specific responses were detected using biotinylated mouse monoclonal antibodies against human IgG1, IgG3, IgG4 (Sigma) and biotin-conjugated mouse monoclonal anti-human IgE Fc (Human Reagent Laboratory, Baltimore, MD) diluted 1:100 in blocking solution for 2 h at RT. Arrays were washed again, and then incubated for 2 h in streptavidin Cy-5 (diluted 1:400 in blocking solution), and washed one final time. Microarrays were scanned using a GenePix microarray scanner (Molecular Devices). The data was analyzed using the "group average" method[69], whereby the mean fluorescence was considered for analysis. Briefly, the fluorescence average of negative control sera was subtracted as the background from the raw data, and then analyses were performed for each protein and immunoglobulin isotype using the average of the negative control as the cut-off for each isotype. Only proteins for which a fluorescence signal above the cut off was detected were considered to be positive.

2. Supplementary Results and Discussion

Genome features

Assembly accuracy and estimation of genome completion

The assessment of the presence of assembly errors was performed using an in-house PERL script which calculated the fraction of the satisfactory constraints from the Newbler's PairStatus file. In summary, 89% of the pairs were satisfactory, of which 9% were flagged as false pairs (where false pairs are defined as those in which both paired ends are mapped, but they are either mapped to the same contig with incorrect orientation or, or the distance between the pairs was outside of the accepted range for the whole genome shotgun library). Feature-response curves (FRC; all-feature, coverage and k-mer based; **Supplementary Fig. 1**) were generated based on methodology outlined in previous studies[70,71].

Three separate methods were used to estimate genome completion, of which two were used to estimate genome size:

a) One of the methods used for estimating the extent of completeness of an assembly is to determine the extent to which core eukaryotic genes (CEGs) can be annotated in the assembly[72,73]. For this study, CEGMA analysis indicated that the genome assembly was 91% complete[72].

b) The genome size was estimated by calculating an estimate for the alignment depth, which was obtained as the peak alignment depth in the histogram provided by Newbler (**Supplementary Fig. 2**) during the assembly process of the whole genome shotgun reads. Based on this estimated depth, the estimated N. americanus genome size is 273.6 Mb, which means that the 244 Mb reported assembly represents 89.2% completion of the genome.

c) For the second genome size estimation (and the third genome completion estimation), the actual depth of reads was calculated using a k-mer frequency distribution analysis. A program based on this approach, GSP (Genome Size Predictor, http://gsizepred.sourceforge.net/), has been used to estimate the genome size of many published genomes[74-76]. Using a low frequency cut-off (parameter 'e') we corrected for sequencing error-based biasing of the k-mer frequency distribution. We used two separate values of this parameter (e=4 and e=5, which is default for the program) to get the estimated genome size values for a range of k-mer lengths (**Supplementary Fig. 3** The genome size estimate saturates at approximately k=18, where the values are 247.8 Mb and 279.8 Mb respectively for the two e values. These values translate to 98.5% and 87.2% completion of the 244 Mb reported assembly, respectively.

Based on these three independent approaches, the reported assembly represents over 90% of the genome and contains at least 92% of the genes encoded in *N. americanu*s, and the estimated *N. americanus* genome size is 273.6 Mb.

Repeat family characterization

In total, 669 repeat families were predicted, including 311 transposable elements, 132 DNA transposons, 98 long terminal repeat (LTR) retrotransposons, and 73 non-LTR retrotransposons (**Supplementary Table 2**). A total of 518 low complexity repeats covering 0.2% of the total bases were also identified. The sequences of all the 669 repeat families have been deposited to nematode.net (www.nematode.net[77]) under the *N. americanus* species hub (http://nematode.net/NN3_frontpage.cgi?navbar_selection=speciestable&subnav_selection=Nector_americanus).

Gene finding and annotation

A total of 19,151 genes were predicted by the following gene prediction algorithms: maker-augustus (7,625), maker-snap (7,437), maker-genesh[14] (1,309), augustus[13] (2,656), snap[12] (98), fgenesh (www.softberry.com) (26). Of the 19,151 genes, 11,229 (58.6%) genes are intact on a single contigs and 7,922 (41.4%) span multiple contigs. The majority of the genes were confirmed by RNA-Seq data (15,825; 82.6%) for two developmental stages (infective L3 - iL3 and adult). The *N. americanus* gene density of 34% falls within the range of other published nematode gene densities (8%-44%), and the genome size (244.1Mb) falls within the range of other published nematode genome sizes (63.5Mb-344.8Mb) as shown below:

Nematode species	Genome Size (Mb)	Number of Scaffolds	Number of Genes	Genes / Mb of genomic DNA	Protein-coding gene density
N. americanus	244.1	11,864	19,151	78.5	34%
H. contortus[78,79]	344.8	20,232	22,705	66.4	8%
A. suum[80]	272.8	29,831	18,542	68.0	44%
T. spiralis[53]	63.5	6,863	16,380	257.8	26%
C. elegans[81]	100.3	7	26,248	261.7	25%
D. immitis[81]	88.3	16,061	12,857	145.6	18%
B. malayi[81]	94.1	9,829	17,989	191.1	14%

A total of 14,211 (74%) *N. americanus* proteins shared homology to proteins from other organisms with available protein sequences (Blastp vs NR, bitscore 35, 55% identity). Functional annotation of the predicted proteins based on primary sequence comparisons was accomplished by identification of Interpro (IPR) domains, Gene Ontology (GO) and KEGG associations [15,21,29]. A total of 4,961 unique domains are identified from 10,975 *N. americanus* genes (**Supplementary Table 5**). Of the 5,806 IPR domains predicted in both *N. americanus* and *C. elegans* (4,961 and 5,463 each, respectively), 1,150 domains had a greater number of genes associated with them in *N. americanus*, 2,430 had the same number of genes associated with

them, and 2,226 had more genes associated with *C. elegans*. GO terms are associated with 8,451 *N. americanus* genes (**Supplementary Table 5**); 736 different 'molecular function' terms are assigned to 7,532 genes, 172 'cellular component' terms to 3,097 genes, and 503 biological processes terms to 5,104 genes of *N. americanus.*

Exon/intron comparisons

The estimated mean gene footprint for genes that do not span a gap is 4,288 bp, and the median exon and intron lengths of these genes (112 and 108 bp, respectively) are comparable with those described for other nematodes[82]. However, the mean exon length in *N. americanus* (130 bp) is significantly less than in *C. elegans* (202 bp) ($p<10^{-10}$), and the mean intron length in *N. americanus* (380 bp) is significantly more than in *C. elegans* (303 bp; $p<10^{-10}$; **Fig. 1a**). An analysis utilizing the orthologous and non-orthologous gene groups showed a statistically significant difference in the exon length of orthologous genes between the two species (127 and 202 bp for *N. americanus* and *C. elegans*, respectively; $p<10^{-10}$), whereas the average intron length of orthologous genes is not statistically different between the two nematodes (358 and 356 bp, respectively; p=0.65). On average, orthologous genes have significantly more introns than non-orthologous genes in both *N. americanus* (6.5 and 3.1 introns per gene, respectively; $p<10^{-10}$) and *C. elegans* (6.4 and 4.1 introns per genes, respectively; $p<10^{-10}$), but the orthologous genes do not have statistically different intron numbers between the species (6.5 and 6.4 introns per gene, respectively; p=0.69). These results show that orthologous genes contain similar average intron length and number between the two nematodes, which is not true for any other comparison made, including stage-specific genes (**Fig. 1b**). Introns in *C. elegans* genes that are orthologous to *N. americanus* genes are significantly longer than introns in non-orthologous

C. elegans genes (**Fig. 1c**), which may indicate a diversity of function for these genes, because longer introns are thought to contain functional elements that are present in addition to what might be regarded as 'normal' intron structure[83]. As expected for eukaryotic organisms, the average length of the first intron is larger than the average length of all of the other introns in orthologous and non-orthologous gene sets in both *N. americanus* and *C. elegans*[83] Mean exon and intron lengths differ between the two species and the two stages in *N. americanus*; significantly longer introns were detected for genes that were orthologous between *N. americanus* and *C. elegans* than in non-orthologous genes, and iL3-overexpressed *N. americanus* genes had longer introns than adult-overexpressed genes (**Fig. 1b**). These longer introns may indicate a greater diversity of regulation for these gene sets[83].

Operons

A combination of two methods was employed to identify putative operons in *N. americanus* genome. First, using the 1,391 inferred operons from the *C. elegans* genome[84] (1356 of which had more than 1 non-pseudogene locus), we predicted 383 putative operons representing 859 genes (4.5% of all genes) in the *N. americanus* genome (**Fig. 1d**), of which 210 (55%) are conserved with *C. elegans* when gene members and orientation are considered (**Supplementary Table 4**). These numbers are significantly higher than those detected in 4 other parasitic species (**Supplementary Table 4**), reflecting the phylogenetic distance among the species (*C. elegans* and *N. americanus* are the only Clade V nematodes in the comparisons). The extremely low number of genes in *T. spiralis* confirms the reported highly divergent and low percent of splice leader sequence identified in *T. spiralis* cDNAs[53]. We also observed some cases with potential inversions in *N. americanus* compared with the *C. elegans* operon structures (see

Supplementary Fig. 6 for an example). The average intergenic region size for conserved operons in *N. americanus* is 9,141 bp, compared with 429 bp in *C. elegans*. Genes within operons had similar expression profiles than random subsets of non-operon genes ($p<0.0001$), indicating that the genes within operons are transcribed together under the same regulatory control, as previously suggested[85,86].

This set of putative operons was expanded using a search for sequences similar to *C. elegans* SL2 sequences[23]. 701 genes were identified that are likely to be trans-spliced with SL2-like sequences, and hence are putative operon genes. A comparison of this set of genes with the ones obtained via *C. elegans* orthology showed a highly significant enrichment of operon genes in the SL2 associated gene set (133 genes in common; $p<2.2x10^{-16}$). Searching for sequences similar to *C. elegans* SL1 led to identification of 2244 SL1 trans-spliced genes, 316 of which were also found to be trans-spliced with SL2-like sequences, and are likely "hybrid operons" with internal promoters[87,88]. While SL1 sequences are known to be trans-spliced to downstream genes in an operon, they are also trans-spliced to non-operon genes, and hence do not verify the presence of an operon. Nevertheless, the SL1 associated gene set is also significantly enriched in the operon genes (166 genes in common; $p=8.1x10^{-9}$). As expected, the genes that are found to be trans-spliced only with SL1 sequence have a lower, though still significant (p-value=$1.1x10^{-5}$), enrichment of operon genes. In all, we confirmed SL trans-splicing of 264 out of the 879 operon genes identified through homology.

The identified number of operons and genes within operons is likely an underestimate of the total number of existing operons and operon genes in *N. americanus* due to i) possible missed detection of SL2 due to sequence variation specific to *N. americanus*; ii) lower detection of transcripts by transcriptome assemblies compared to RNAseq reads, resulting in missed 5' end of

the genes; iii) the draft nature of the *N. americanus* genome providing suboptimal analysis due to existence of gapped genes and limited ability to detect gene orders, especially if the operon genes are on different supercontigs.

Transcriptional differences between infective and parasitic stages

Secretome and degradome

Excretory/secretory (E/S) proteins are present at the host-parasite interface and play roles in parasitic infection and host immune system regulation. *N. americanus* encodes 1,590 genes (8.3% of all genes) with translations predicted to have signal peptide for secretion, comparable to other parasitic nematodes (*T. spiralis*, 6.8%; *A. suum*, 8%) but less than *C. elegans* (14.9%) and 4,785 genes secreted through the non-classical pathways (**Supplementary Table 7**).

Proteases and protease inhibitors are expressed and secreted by *N. americanus* larvae to invade and penetrate host skin (in order to migrate through and feed on tissues), by adults to inhibit the coagulation of the blood meal and enable feeding and digestion, and by both stages to evade host immune responses[89]. Serine and metalloproteases have previously been identified in secretory products of nematode larvae[90]. In *A. caninum*, evidence indicates that an astacin-like metalloprotease secreted by infective larvae may have a role in tissue-migration[91]. In addition, adults of the canine hookworm *Ancylostoma caninum* secrete mediators, such as fatty acid/retinol binding protein[92] and tissue inhibitors of metalloproteases[93], are also detected in *N. americanus*.

Overall, nematode proteases can function in the intestine locally for nutrient acquisition or can interact more broadly with the host. As the first step in nutrient acquisition in the blood, single-domain serine protease inhibitors are secreted which function in the inhibition of

coagulation, which is critical for consuming blood from intestinal capillaries. The vast majority (87 of 107; 81%) of endogenous protease inhibitors in *N. americanus* belong to a group of serine protease inhibitors. Given that, in humans, serine proteases are extensively involved in diverse physiological functions (including blood coagulation, immune responses and digestion), we reason that the occurrence of diverse serine protease inhibitors in *N. americanus* is critical for the parasite's survival in the host environment. The number of anti-coagulants encoded by different species of *Ancylostoma* varies; while *A. caninum* encodes 6 anticoagulants that inhibit coagulation factor Xa or XIa or the VIIa-tissue factor complex, *A. ceylanicum* encodes a single anticoagulant that inhibits Xa and VIIa-tissue factor[94].

A cascade of proteases is also necessary to facilitate hemoglobin degradation and nutrient acquisition. *N. americanus* lyse and catabolize red blood cells using pathways consisting of hemolytic and proteolytic enzymes. Current evidence shows that hemoglobin is initially digested by aspartic proteases, followed by the degradation of smaller peptides by cysteine proteases, and then metalloproteases[95,96]. These three classes of proteases are highly represented in *N. americanus*.

The *N. americanus* genome encodes three aspartic protease families with 44 genes in total, of which pepsin (A01) is the most abundant (86%), similar to other nematodes included in this study except for *T. spirialis* and *B. malayi*, which encode a member of the retropepsin family (A2). Among the proteases, aspartic proteases have the largest proportion of proteins with signal peptides for secretion (32% compared with 12-15% for all other protease families, **Supplementary Table 7**). The degradome of N. americanus contains a significantly higher proportion of aspartic proteases than its human host ($p < 10^{-15}$, according to a FDR-corrected binomial distribution test with a $p \leq 10^{-5}$ theshold), but contains a similar proportion compared to

other nematodes (**Supplementary Fig. 9a**). Eighteen (18) of the aspartic proteases are overexpressed in adult compared with just four in iL3 (**Supplementary Fig. 9b**). A group of aspartic proteases found in *A. caninum* are also found in *N. americanus*, including Na-APR-1[97,98].

Of the 105 cysteine protease genes encoded in the *N. americanus* genome, 20 are adult-overexpressed and just one is iL3-overexpressed. The most abundant cysteine proteases are from the papain family (C01; 26 genes) and the ubiquitin-specific protease family (C19; 24 genes). Cathepsin B cysteine proteases (members of the C01 family) have been found to be over-expressed in the gut of adult *N. americanus*, suggesting a role in the digestion of host proteins to obtain nutrients[99]. Cysteine proteases (in particular the papain family, C01) are strongly enriched among adult-overexpressed genes ($p=9.1*10^{-6}$, compared to p=0.1 for iL3) and have been reported as vaccine targets in parasitic nematodes[100-102]. Proportionally, *N. americanus* does not encode significantly more or fewer cysteine proteases than other nematodes or outgroups.

N. americanus also encodes 246 metalloprotease genes which belong to 21 different families; however, 75% of them belong only to the 5 most abundant families (>10 genes per family). Metalloproteases are enriched among adult-overexpressed genes ($p=1.1*10^{-4}$), indicating a significant role for them during invasion of and establishment in the host. *N. americanus* contains significantly more metalloproteases (proportionally) than *C. elegans* and *T. spiralis* ($p = 3.5x10^{-7}$ and $3.3x10^{-8}$, respectively), significantly more than the flatworms *C. sinensis* and *S. haematobium* ($p < 10^{-15}$ and $p = 1.0x10^{-13}$, respectively), and significantly fewer than their human host ($p = 2.1x10^{-6}$).

Overall, the *N. americanus* genome encodes 592 proteases (~3% of the predicted proteome, **Supplementary Table 7**), which is more than any other nematodes in comparative

studies. A total of 47 proteases are over-expressed in iL3s, including 5 astacin metalloproteases which in the canine hookworm *A. caninum* are shown to be involved in tissue penetration[91]. Although there are significant differences in the quantities of some specific proteases compared to other species, some members of the proteolytic system of *N. americanus* are similar to other organisms[103], and the composition of proteases correlates with the corresponding inhibitors. Studying proteases and their targets in more detail in *N. americanus* and making comparisons to other closely and distantly related species should provide insight into the parasite's interplay with the host. Additional details of proteases classes and membership in nematodes and in Platyhelminthes species are available on www.nematode.net[104].

Protease inhibitors

Endogenous protease inhibitors regulate protease activity[103]. Serine protease inhibitors (SPIs) protect adult hookworms from the digestive environment within the host. Serin proteases are prominent in the small intestine[105-107], and include trypsin, chymotrysin and elastase, which can mediate hookworm-associated growth delay of the host[108]. In *N. americanus*, endogenous serine protease inhibitors appear to be the most abundant protease inhibitors (87 of 107, 81%). A comparative investigation showed that the abundance of serine protease inhibitors and metalloproteases was consistent for all available nematode genomes, including that of *C. elegans*.

We found that "Serine-type endopeptidase inhibitor activity" (GO:0004867) was significantly enriched among genes overexpressed in the adult stage ($p=1.6x10^{-4}$), with 21 of 47 (44.7%) of the adult-overexpressed protease inhibitors belonging to this category (13 kunitz-type, 7 serpin I4-type and 1 kazal-type)[32]. To date, the most characterized SPIs in hookworms are

kunitz-type. However, the present data show that multiple types of serine protease inhibitors may contribute to the survival of adult *N. americanus* in the gut. Studying these molecules in detail will provide insight on the parasite's interplay with the host.

Cysteine protease inhibitors (cystatins) released by helminths are also believed to modulate host immune and inflammatory responses[109]. Cysteine proteases are involved in the degradation of protein to antigenic peptides, and in two important processes of antigen presentation[110]. Nematode cystatins inhibit the mammalian lysosomal cysteine proteases cathepsins L, S, and B[111-113]. Since these proteases are used to process antigens for presentation by the major histocompatability complex (MHC) class II APCs and digest the MHC II-associated invariant chain chaperone prior to peptide antigen loading, the inhibition of cathepsins L and S most likely allows the parasite to evade host immune responses. Members of the cystatin family (I25) of cysteine protease inhibitors are found in *N. americanus* and other parasitic nematodes (except *T. spiralis*).

Proteomic analysis of somatic worm extract

A mass spectrometry-based proteomics analysis was performed using whole adult *N. americanus* worms (Full Methods Online). Proteins with at least two significant ($p < 0.05$) peptides (corresponding to a 0.0% protein false discovery rate) were considered detected, and corresponded to 458 genes in the *N. americanus* genome. These genes had relatively high expression levels in the adult stage (**Supplementary Fig. 10a**), with 143/458 (31.2%) of them being overexpressed in the adult stage ($p<10^{-15}$), and only one of them being overexpressed in the iL3 stage ($p=7.1x10^{-7}$ for depletion; **Supplementary Table 7**). The level of protein detection from the adult worms (measured by the number of peptides) correlated with the expression level

of their corresponding genes on a log scale (Pearson correlation = 0.44, p = $5*10^{-23}$; **Supplementary Fig. 10b**). The genes encoding the proteins detected were enriched for proteases ($p=4.9x10^{-7}$), SPIs ($p=1.8x10^{-4}$), as well as signal peptides ($p=4.7x10^{-11}$) and non-classical secretion sequences ($p=7.0x10^{-13}$), and were depleted for transmembrane domains ($p=6.6x10^{-13}$). According to FUNC[114] GO enrichment testing, 43 terms were significantly enriched ($p<=0.01$), and 25 terms were significantly depleted among the genes supported by proteomics in this analysis (**Supplementary Table 6**).

Transcription factors and transcription factors binding sites

Regulation of gene expression and control of development and homeostasis in *N. americanus* is accomplished by at least 217 transcription factors (TFs), which is fewer compared to the TFs encoded by *C. elegans* (**Supplementary Table 3**). This difference may partially relate to an underestimation of the number of TFs in *N. americanus* due to the draft-nature of the genome, or it may be due to the parasitic life style of *N. americanus* which may require less unique pathways due to the more confined host environment and the reliance on host proteins and compounds. There are 43 TFs which do not have orthologs in *C. elegans*, suggesting that *N. americanus* has evolved some unique gene regulatory mechanisms. Of the 186 TFs which were transcriptionally detected, 28.5% (53) were differentially expressed, of which the majority (48, 91%) are iL3-overexpressed. Among these iL3-overexpressed TFs are the fork head transcription factors that are involved in promoting dauer in *C. elegans* and the dog hookworm (*A. caninum*). Post-translational regulatory mechanisms (phosphorylation-dephosphorylation cycles) have already been suggested for these transcription factors[115-117]. The most strongly adult-enriched TF encodes a cold-shock domain (CSD) protein (NECAME_08819), a member of a class of stress-related

TFs whose role is well defined in bacteria, plants and mammals[118] but not in nematodes. Given the complex nature of the migration of developing hookworms, CSD proteins/cold-shock responses may be critical for *N. americanus*.

TFs operate via their binding to the specific DNA motifs, i.e. binding sites that are in the proximal regions near the transcription start site[119-121]. The predicted binding sites in the *N. americanus* genome for all the differentially expressed genes are associated with 19 iL3-overexpressed and 1 adult-overexpressed TF. Although TF binding sites have yet to be precisely defined, the present prediction of TFs defined the first regulatory network for *N. americanus*, and further investigations of the TFs and their gene targets will aid in deciphering the regulation of gene expression in this hookworm.

Pathogenesis and immunobiology of hookworm disease

Some potential molecular mechanisms for successful human parasitism by *N. americanus* were explored by comparing its genome with previously characterized immune-pathogenesis/host-parasite interaction-related genes. Orthologous groups for *N. americanus* proteins were built with 4 parasitic nematodes, 1 non-parasitic nematode, 4 flatworms, the human host and 2 outgroups (**Supplementary Table 8**).

SCP/TAPS proteins

The immune-pathogenesis/host-parasite interaction-related protein-coding genes include SCP/Tpx-1/Ag5/PR-1/Sc7 (SCP/TAPS) proteins, neutrophil inhibitor factor (NIF), hookworm platelet inhibitor (HPI), transforming growth factor beta (TGF-β), cysteine protease inhibitors

(cystatins), galectins, C-type lectins (C-TL), peroxiredoxins (PRX) and glutathione S-transferases (GST).

The SCP/TAPS protein family belongs to the cysteine-rich secretory protein (CRISP) superfamily with immunomodulatory activity[47,122], and have been identified in the ES products of *N. americanus* and *A. caninum*[123] as well as the murine strongylid nematode, *Heligmosomoides polygyrus*[124]. The proteins contain single or double CAP domain(s) and may also include non-CAP domain(s) with variable functions and inter-domain length (up to 170 amino acids). There has been limited investigation into hookworm SCP/TAPS proteins since they were proposed to have fundamental roles in the host-parasite interplay over a decade ago[43]. The only exception to this is Na-ASP-2, which has been the subject of both structural studies as well as vaccine clinical trials[125,126].

In *N. americanus*, SCP/TAPS gene members are strongly over-represented compared to other nematodes, including *C. elegans*, *B. malayi*, *L. Loa*, *A. suum* and *T. spiralis* (which possess 34, 10, 12, 20 and 15 genes, respectively, compared to 137 in *N. americanus*), suggesting that they might be involved in multiple (and possibly distinct) aspects of hookworm biology. Interestingly, those proteins are absent from the vertebrate host of *N. americanus* based on orthologous group studies (**Supplementary Table 8**)[127]. Of the 137 CAP-domain/Allergen V5/Tpx-1 genes, 96 are restricted to *N. americanus*, which further suggests that they perform functions specific to this parasite.

The *N. americanus* SCP/TAPS proteins contain either single CAP domains (107) or double CAP domains (30). Additional structural units were also detected and may deliver alternate functionalities. SCP/TAPS protein trees constructed using two methods (Bayesian inference and Neighbor Joining of the full-length sequence) have similar topology (**Fig. 3c**,

Supplementary Fig. 12). Both suggested that these proteins in nematodes have originated prior to parasitism, and that some of them greatly expanded after hookworm speciation. In some cases, overlapping of the domains was observed, which may be due to gaps in the scaffolds resulting from the draft nature of the genome. Phylogenetic analysis of *N. americanus* SCP/TAPS superfamily members, which includes 61 sequences derived from ungapped scaffolds longer than 124 amino acids (the recorded length of CAP domain in Pfam database) showed significant diversity within *Necator* SCP/TAPS proteins (**Fig. 3c**). At least three major groups were categorized. The presence of one group with a limited repertoire of orthologs in *C. elegans* suggests that nematode SCP/TAPS proteins may have originated prior to parasitism. In addition to the structural differences, far fewer members of *C. elegans* SCP/TAPS proteins were found to be developmentally regulated between the iL3 ("dauer") stage and the adult stage compared to *N. americanus*, which also suggests critical roles for these proteins in parasitism. The dataset described in this study provides a valuable resource for further investigation of this group of proteins, and provides structural as well as transcriptional information to narrow down potential targets of interest.

Other immunomodulators

C-type lectins (C-TLs) are a family of carbohydrate-binding proteins proposed to be involved in immune invasion, and possibly other host-parasite interactions. C-TLs are commonly found in nematodes and mammals, and are believed to interrupt anti-parasite immune responses or interfere with host blood clotting[128]. Parasite C-TLs may bind to nematode carbohydrates and mask them from the host immune cells[129]. There are 57 *N. americanus* C-type lectin-like genes (IPR016186), 11 of which were adult-overexressed, and one of which was iL3-overexpressed.

One CTL was also previously characterized as being exclusive to the adult *N. americanus*[130]. In *N. americanus*, eight genes encode galectins, a family of lectins that bind to various β-galactoside-containing glycans. Galectins are found in *A. caninum* ES products[35], and in *H. contortus*, a galectin-like protein is an attractant for eosinophils[131]. Other hookworm proteins may also regulate immune responses, based on the fucntions of orthologous proteins in other parasitic nematode ES products, including calreticulin[132], transthyretin-like (TTL) proteins[35,133,134], peroxiredoxins[135], and glutathione S-transferases (GST)[136,137]. Some of these, such as the hookworm GSTs[125,138,139], have been extensively studied as promising candidates for developing recombinant hookworm vaccines. Many other proteins with putative immuno-modulatory and immuno-evasive functions have also been discovered (**Supplementary Table 5**).

Prospects for new interventions

Ivermectin targets

We compared the six *C. elegans* glutamate-gated chloride channel genes: *avr 14/15* and *glc 1-4*[140], which clustered into two ortholog groups with three *N. americanus* genes. Sequence alignments of all the orthologs suggest that the differences in effectiveness may arise from sequence variations at the binding region of ivermectin. In the crystal structure of *C. elegans* GluCl channel (*glc-1*) in complex with ivermectin[50], ivermectin inserts deeply into subunit interfaces on the periphery of the transmembrane domains and makes important contacts with the M2 (+) pore-lining α helix and the M2–M3 loop, with a critical residue (Ser 260) forming a hydrogen bond with the deeply buried cyclohexene ring of ivermectin (**Supplementary Fig.**

13a). A serine residue at this position is correlated with direct activation by ivermectin in other Cys-loop receptors. Glycine and $GABA_A$ receptors have a serine in the equivalent position and are directly activated by ivermectin[141]. The sequence alignment reveals that for all the *N. americanus* genes, the equivalent residue to Ser 260 is either alanine or glutamine (**Supplementary Fig. 13b**), which leads to the putative disruption of the hydrogen bond and the loss of the anchoring role for the binding of ivermectin (**Supplementary Fig. 13c-d).** In *C. elegans*, the essential serine residue is present in at least two of the six GluCl channel genes, while the lack of a clear ortholog of the ivermectin-sensitive genes within *N. americanus* may explain its relative insensitivity to the drug.

Nuclear receptors

Nuclear receptors (NRs) are transcription factors which are typically regulated by lipophilic molecules to control development, metabolism and homeostasis[142], and are promising anthelmintic targets[143]. The *N. americanus* genome encodes 115 NR genes (**Supplementary Table 9**) based on BLAST searches against the *C. elegans* proteome, which contains 284 NR proteins[144] (cutoff 1E-10). NRs are only found in metazoans, and are more prevalent in nematodes compared with *Drosophila* (21) and human (48)[145]. Most NRs in *C. elegans* have not been studied in detail, but characterization of *N. americanus* NRs will enable the elucidation of transcription hierarchies and connectivity with other signal transduction pathways.

Neuropeptides

Neuropeptides have been postulated to play important roles in neuronal signal processing and behavior modulation in *C. elegans*[146], and neuropeptides and their putative receptors may serve

as potential therapeutic targets. In the *N. americanus* genome, we have annotated 68 neuropeptide genes orthologous to known *C. elegans* neuropeptides (**Supplementary Table 9**), including 16 genes coding FMRFamide-related peptides (FLP) peptides, 50 genes coding non-insulin, non-FMRFamide-related neuropeptides (NLP) peptides and 2 genes coding insulin-like peptides (INS) peptides. The FLP and NLP peptides are linked with intracellular communication and motor functions, which has been successfully targeted by anthelmintics[147].

Metabolic chokepoints

A metabolic pathway chokepoint is an enzyme that produces or consumes a unique compound[148]. Inhibition of a chokepoint blocks the entire pathway, leading to the accumulation of a unique substrate or the starvation of a unique product, making them an attractive therapeutic target. *N. americanus* has 938 metabolic pathway enzymes, with 318 (34%) classified as a chokepoint. The 318 *N. americanus* chokepoints were prioritized as potential therapeutic targets, based on conservation among 12 other species (nematodes, flatworms, and an outgroup) used in the comparative analysis, display of a lethal RNAi phenotype, expression in parasitic stages of nematodes, expression in particular tissues, sequence dissimilarity to human, presence in pathways, and if they are an expression bottleneck (**Supplementary Table 12**). An expression bottleneck was defined as a rate-limiting step of a reaction cascade and determines the overall rate of formation of the product. Targeting a bottleneck enzyme will more likely reduce a pathway module's output compared to other enzymes that may be present in relative excess.

It should be noted that any study where RNA abundance levels are used as a proxy for corresponding protein levels and/or activity should be interpreted with caution, as there are many potential mechanisms which may modify the correlation between RNA abundance on protein

activity. These include translation efficiency, protein modification and degradation dynamics, and enzyme kinetics. Studies investigating the correlation between RNA levels and protein levels have found an overall positive correlation, although the correlation levels have varied widely, especially between different sets of proteins[149-153]. Given the difficulty of obtaining large-scale proteomic and enzyme kinetics data, RNA expression levels have extensively been used to obtain important information about the biology of organisms. Here, RNA expression levels will be used to help define bottlenecks in this analysis, but in the future, these types of analyses may benefit from high-throughput proteomics which cover the majority of the expressed proteins, and enzyme kinetics information.

A total of 40 adult-stage bottlenecks were identified using RNA-Seq data to assess the abundance of the 37 potentially complete metabolic modules, 36 of which are also potentially complete in the human proteome which has 59 complete metabolic modules (**Supplementary Table 13**).

Dihydropyrimidinase (DHP; also known as hydantoinase), the highest scored chokepoint, is a hydrolase which catalyzes the second step of the reductive pyrimidine degradation (the reversible hydrolytic ring opening of dihydropyrimidines)[154]. The two substrates of DHP are 5,6-dihydrourarcil and water, and its product is 3-ureidopropanoate. DHP is expressed in the hypodermis of *C. elegans* during the early larva stage, in a nerve cell, and in the body wall muscle throughout all stages[155]. DHP has also been identified as a tumor suppresor target[156].

The third highest-scoring chokepoint is 5'-nucleotidase, an enzyme that catalyzes hydrolysis of a nucleotide into a nucleoside and a phosphate. Nucleotides are signaling molecules that are secreted by the host, and may be involved in immune responses[157]. *T. spiralis* uses 5'-nucleotidase to catalyze the degradation of extracellular nucleotides, which may have a

role in regulating host purinergic signaling during infection. A bacterial pathogen, *Burkholderia cepacia*, secretes more 5'-nucleotidase in clinical strains compared to environmental strains[158], and may play an important role in evading host defense. 5'-nucleotidase can be inhibited by plant compounds lycorine and candimine in *Trichomonas* vaginalis[158,159].

The fourth highest-scoring chokepoint is Nucleoside-diphosphate kinase (NDK) which catalyzes the exchange of phosphate groups between different nucleoside diphosphates. NDK is secreted in *T. spiralis, H. contortus* and *T. circumcinta,* and has been suggested to play a role in modulating host responses[160-163].

Prioritization of compounds that target chokepoints

Prioritization of chokepoint inhibitors revealed 6 compounds with a score of 7 (which is the highest score possible; see Full Methods Online; **Supplementary Table 14**): pyrazinamide (DB00339) targeting fatty-acid synthase (EC 2.3.1.85), perhexiline (DB01074) targeting carnitine O-palmitoyltransferase (EC 2.3.1.21), propylthiouracil (DB00550) targeting iodide peroxidase (EC 1.11.1.8), carbidopa (DB00190) targeting aromatic-L-amino-acid decarboxylase (EC 4.1.1.28), azathioprine (DB00993) targeting adenylosuccinate lyase (EC 4.3.2.2), and azelaic acid (DB00548) targeting 3-oxo-5-alpha-steroid 4-dehydrogenase (EC 1.3.99.5). Existing drugs for repositioning can be directly tested in animal models for efficacy, and also optimized for affinity, specificity, and pharmacological properties.

Pyrazinamide (DB00339) is a fatty-acid synthase inhibitor and an approved antibiotic drug used to treat tuberculosis[164]. Fatty-acid synthase (FAS) catalyzes fatty acid biosynthesis by catalyzing the synthesis of palmitate from acetyl-CoA and malonyl-CoA in the presence of

NADPH[165]. Fatty acids are necessary for energy production and storage, cellular structure, and also serve as intermediates in the biosynthesis of other biologically important molecules.

Perhexiline (DB01074) is a carnitine O-palmitoyltransferase (CPT) inhibitor and a prophylactic antianginal agent[166]. CPT is a mitochondrial transferase enzyme involved in the metabolism of palmitoylcarnitine into palmitoyl-CoA. CPT catalyzes an essential step in the beta-oxidation of long chain fatty acids, and has been exploited for treating various human diseases[167].

Propylthiouracil (DB00550) is an iodide/thyroid peroxidase inhibitor used to treat hyperthyroidism by decreasing the amount of thyroid hormone produced by the thyroid gland[168]. Thyroid peroxidase is an enzyme used in the thyroid that adds iodine onto tyrosine residues on thyroglobulin for the production of thyroid hormones[169].

Carbidopa (DB00190) is an aromatic-L-amino-acid decarboxylase (AAAD) inhibitor and an approved drug used to treat Parkinson Disease[170]. AAAD is a lyase that catalyzes several decarboxylation reactions: L-DOPA to dopamine, 5-HTP to serotonin, and tryptophan to tryptamine[171].

Azathioprine (DB00993) is an adenylosuccinate lyase (ASL) inhibitor and an approved drug to treat inflammatory bowel diseases such as Crohn Disease and autoimmune diseases such as rheumatoid arthritis[172,173]. ASL catalyzes two reactions in the de novo purine biosynthesis pathway: adenylosuccinate into adenosine monophosphate (AMP) and fumarate, and 5-aminoimidazole- (N-succinylocarboxamide) ribotide (SAICAR) into 5-aminoimidazole-4-carboxamide ribotide (AICAR) and fumarate[174]. ASL is critical for survival since it is involved in creating purines (which are needed for cellular replication) and because it helps to regulate a large number of metabolic processes by controlling the levels of AMP and fumarate in the cell.

Azelaic acid is a 3-oxo-5-alpha-steroid 4-dehydrogenase inhibitor and an approved antibacterial drug used for treating acne[175]. 3-oxo-5-alpha-steroid 4-dehydrogenase, also known as 5α-reductase, participates in three metabolic pathways: bile acid biosynthesis, androgen and estrogen metabolism, and prostate cancer[176]. 5α-reductase catalyzes 3-oxo-5α-steroid and an acceptor into 3-oxo-Delta4-steroid and a reduced acceptor.

A platform for post-genomic explorations – the N. americanus immunome

As an example of a post-genomic application, using the knowledge gained on secreted proteins, RNA-Seq based expression and using high throughput cell-free protein expression[177], we developed a protein microarray containing approximately 25% of the *N. americanus* proteins with predicted signal peptide for secretion (620 proteins). Screening the resultant protein array with sera from individuals resident in a hookworm-endemic area of Brazil showed distinct antigen-recognition profiles related to the intensity of infection. For example, a relatively small secreted protein (NECAME_12931) was recognized by all infected individuals (**Supplementary Fig. 19**) and has a weak similarity to hypothetical proteins from *Caenorhabditis sp*. but did not appear to have homologues in parasitic nematodes, indicating that these antigens might form the basis of sensitive and specific serodiagnostic tests. This approach highlights the utility of the genome information in exploring the immuno-biology of human hookworm disease and accelerating antigen discovery for the development of vaccines and diagnostics.

References

1. Wylie, T. et al. NemaPath: online exploration of KEGG-based metabolic pathways for nematodes. *BMC Genomics* **9**, 1471-2164 (2008).
2. Eswar, N. et al. Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci* **2**(2007).
3. Morris, G.M. et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry* **30**, 2785-91 (2009).
4. Jian, X. et al. Necator americanus: maintenance through one hundred generations in golden hamsters (Mesocricetus auratus). I. Host sex-associated differences in hookworm burden and fecundity. *Exp Parasitol* **104**, 62-6 (2003).
5. Xiao, S. et al. The evaluation of recombinant hookworm antigens as vaccines in hamsters (Mesocricetus auratus) challenged with human hookworm, Necator americanus. *Experimental parasitology* **118**, 32-40 (2008).
6. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-80 (2005).
7. Kohany, O., Gentles, A.J., Hankus, L. & Jurka, J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC bioinformatics* **7**, 474 (2006).
8. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**, 462-7 (2005).
9. Lagesen, K. et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research* **35**, 3100-8 (2007).
10. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* **25**, 955-64 (1997).
11. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S.R. Rfam: an RNA family database. *Nucleic acids research* **31**, 439-41 (2003).
12. Korf, I. Gene finding in novel genomes. *BMC bioinformatics* **5**, 59 (2004).
13. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-44 (2008).
14. Cantarel, B.L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research* **18**, 188-96 (2008).
15. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).
16. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research* **35**, W182-5 (2007).
17. Quevillon, E. et al. InterProScan: protein domains identifier. *Nucleic acids research* **33**, W116-20 (2005).
18. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research* **31**, 365-70 (2003).
19. Finn, R.D. et al. Pfam: clans, web tools and services. *Nucleic acids research* **34**, D247-51 (2006).
20. Marchler-Bauer, A. et al. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic acids research* **39**, D225-9 (2011).

21. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* **40**, D109-14 (2012).
22. Li, L., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* **13**, 2178-89 (2003).
23. Guiliano, D.B. & Blaxter, M.L. Operon conservation and the evolution of trans-splicing in the phylum Nematoda. *PLoS Genet* **2**, e198 (2006).
24. Wang, Z. et al. Characterizing Ancylostoma caninum transcriptome and exploring nematode parasitic adaptation. *BMC Genomics* **11**, 307 (2010).
25. Hancock, J.M. & Armstrong, J.S. SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. *Comput Appl Biosci* **10**, 67-70 (1994).
26. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
27. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-40 (2010).
28. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).
29. Hunter, S. et al. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic acids research* **40**, D306-12 (2012).
30. Kall, L., Krogh, A. & Sonnhammer, E.L. A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology* **338**, 1027-36 (2004).
31. Bendtsen, J.D., Jensen, L.J., Blom, N., Von Heijne, G. & Brunak, S. Feature-based prediction of non-classical and leaderless protein secretion. *Protein engineering, design & selection : PEDS* **17**, 349-56 (2004).
32. Rawlings, N.D., Barrett, A.J. & Bateman, A. MEROPS: the peptidase database. *Nucleic acids research* **38**, D227-33 (2010).
33. Prufer, K. et al. FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC bioinformatics* **8**, 41 (2007).
34. Binns, D. et al. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25**, 3045-6 (2009).
35. Mulvenna, J. et al. Proteomics analysis of the excretory/secretory component of the blood-feeding stage of the hookworm, Ancylostoma caninum. *Molecular & cellular proteomics : MCP* **8**, 109-21 (2009).
36. Craig, R. & Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466-7 (2004).
37. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**, 207-14 (2007).
38. Yang, P. et al. Improving X!Tandem on peptide identification from mass spectrometry by self-boosted Percolator. *IEEE/ACM Trans Comput Biol Bioinform* **9**, 1273-80 (2012).
39. Nesvizhskii, A.I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75**, 4646-58 (2003).
40. Matys, V. et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research* **34**, D108-10 (2006).

41. Bryne, J.C. et al. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic acids research* **36**, 15 (2008).
42. 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).
43. Cantacessi, C. & Gasser, R.B. SCP/TAPS proteins in helminths--where to from now? *Mol Cell Probes* **26**, 54-9 (2012).
44. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS computational biology* **7**, e1002195 (2011).
45. Ronquist, F. & Huelsenbeck, J.P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-4 (2003).
46. Larkin, M.A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-8 (2007).
47. Cantacessi, C. et al. A portrait of the "SCP/TAPS" proteins of eukaryotes--developing a framework for fundamental research and biotechnological outcomes. *Biotechnology advances* **27**, 376-88 (2009).
48. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127-8 (2007).
49. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-7 (2004).
50. Hibbs, R.E. & Gouaux, E. Principles of activation and permeation in an anion-selective Cys-loop receptor. *Nature* **474**, 54-60 (2011).
51. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. & Barton, G.J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-91 (2009).
52. Miranda-Saavedra, D. & Barton, G.J. Classification and functional annotation of eukaryotic protein kinases. *Proteins* **68**, 893-914 (2007).
53. Mitreva, M. et al. The draft genome of the parasitic nematode Trichinella spiralis. *Nat Genet* **43**, 228-35 (2011).
54. Punta, M. et al. The Pfam protein families database. *Nucleic acids research* **40**, D290-301 (2012).
55. Lipinski, C.A., Lombardo, F., Dominy, B.W. & Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **46**, 3-26 (2001).
56. Taylor, C.M. et al. Using Existing Drugs as Leads for Broad Spectrum Anthelmintics Targeting Protein Kinases. *PLoS Pathog* **9**, e1003149 (2013).
57. Kumar, S. et al. Mining predicted essential genes of Brugia malayi for nematode drug targets. *PLoS One* **2**, e1189 (2007).
58. Yook, K. et al. WormBase 2012: more genomes, more data, new website. *Nucleic Acids Res* **40**, D735-41 (2012).
59. Hunter, S. et al. InterPro: the integrative protein signature database. *Nucleic acids research* **37**, D211-5 (2009).
60. Hillier, L.W. et al. Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. *Genome research* **19**, 657-66 (2009).
61. Abubucker, S., Martin, J., Taylor, C.M. & Mitreva, M. HelmCoP: an online resource for helminth functional genomics and drug and vaccine targets prioritization. *PLoS One* **6**, e21832 (2011).

62. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research* **38**, 30 (2010).
63. Notredame, C., Higgins, D.G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology* **302**, 205-17 (2000).
64. Sali, A. & Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology* **234**, 779-815 (1993).
65. Knox, C. et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* **39**, D1035-41 (2011).
66. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-504 (2003).
67. Driguez, P., Doolan, D.L., Loukas, A., Felgner, P.L. & McManus, D.P. Schistosomiasis vaccine discovery using immunomics. *Parasit Vectors* **3**, 4 (2010).
68. Davies, D.H. et al. Profiling the humoral immune response to infection by using proteome microarrays: high-throughput vaccine and diagnostic antigen discovery. *Proc Natl Acad Sci U S A* **102**, 547-52 (2005).
69. Sundaresh, S. et al. Identification of humoral immune responses in protein microarrays using DNA microarray data analysis techniques. *Bioinformatics* **22**, 1760-6 (2006).
70. Narzisi, G. & Mishra, B. Comparing de novo genome assembly: the long and short of it. *PLoS One* **6**, 0019175 (2011).
71. Vezzi, F., Narzisi, G. & Mishra, B. Feature-by-feature--evaluating de novo sequence assembly. *PLoS One* **7**, 3 (2012).
72. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-7 (2007).
73. Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. *Nucleic acids research* **37**, 289-97 (2009).
74. Li, R. et al. The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311-7 (2010).
75. Fan, L. et al. Draft genome sequence of the marine Streptomyces sp. strain PP-C42, isolated from the Baltic Sea. *J Bacteriol* **193**, 3691-2 (2011).
76. Li, Z. et al. Genome sequence of the tobacco bacterial wilt pathogen Ralstonia solanacearum. *J Bacteriol* **193**, 6088-9 (2011).
77. Martin, J., Abubucker, S., Heizer, E., Taylor, C.M. & Mitreva, M. Nematode.net update 2011: addition of data sets and tools featuring next-generation sequencing data. *Nucleic Acids Res* **40**, D720-8 (2012).
78. Schwarz, E.M. et al. The genome and developmental transcriptome of the strongylid nematode Haemonchus contortus. *Genome Biol* **14**(2013).
79. Laing, R. et al. The genome and transcriptome of Haemonchus contortus, a key model parasite for drug and vaccine discovery. *Genome Biol* **14**(2013).
80. Jex, A.R. et al. Ascaris suum draft genome. *Nature* **479**, 529-533 (2011).
81. Godel, C. et al. The genome of the heartworm, Dirofilaria immitis, reveals drug and vaccine targets. *FASEB J* **26**, 4650-61 (2012).
82. Sommer, R.J. & Streit, A. Comparative genetics and genomics of nematodes: genome structure, development, and lifestyle. *Annu Rev Genet* **45**, 1-20 (2011).
83. Bradnam, K.R. & Korf, I. Longer first introns are a general property of eukaryotic gene structure. *PLoS One* **3**, e3093 (2008).

84. Blumenthal, T. et al. A global analysis of Caenorhabditis elegans operons. *Nature* **417**, 851-4 (2002).
85. Lercher, M.J., Blumenthal, T. & Hurst, L.D. Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. *Genome research* **13**, 238-43 (2003).
86. Zaslaver, A., Baugh, L.R. & Sternberg, P.W. Metazoan operons accelerate recovery from growth-arrested states. *Cell* **145**, 981-92 (2011).
87. Allen, M.A., Hillier, L.W., Waterston, R.H. & Blumenthal, T. A global analysis of C. elegans trans-splicing. *Genome research* **21**, 255-64 (2011).
88. Huang, P. et al. Identification and analysis of internal promoters in Caenorhabditis elegans operons. *Genome Res* **17**, 1478-85 (2007).
89. Kumar, S. & Pritchard, D.I. Secretion of metalloproteases by living infective larvae of Necator americanus. *The Journal of parasitology* **78**, 917-9 (1992).
90. Tort, J., Brindley, P.J., Knox, D., Wolfe, K.H. & Dalton, J.P. Proteinases and associated genes of parasitic helminths. *Advances in parasitology* **43**, 161-266 (1999).
91. Williamson, A.L. et al. Ancylostoma caninum MTP-1, an astacin-like metalloprotease secreted by infective hookworm larvae, is involved in tissue migration. *Infection and immunity* **74**, 961-7 (2006).
92. 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. *Molecular and biochemical parasitology* **126**, 63-71 (2003).
93. Zhan, B. et al. Molecular cloning and purification of Ac-TMP, a developmentally regulated putative tissue inhibitor of metalloprotease released in relative abundance by adult Ancylostoma hookworms. *The American journal of tropical medicine and hygiene* **66**, 238-44 (2002).
94. Li, D. et al. Identification of an anticoagulant peptide that inhibits both fXIa and fVIIa/tissue factor from the blood-feeding nematode Ancylostoma caninum. *Biochemical and biophysical research communications* **392**, 155-9 (2010).
95. Williamson, A.L., Brindley, P.J., Knox, D.P., Hotez, P.J. & Loukas, A. Digestive proteases of blood-feeding nematodes. *Trends in parasitology* **19**, 417-23 (2003).
96. Knox, D. Proteases in blood-feeding nematodes and their potential as vaccine candidates. *Advances in experimental medicine and biology* **712**, 155-76 (2011).
97. Hotez, P.J. et al. Effect of vaccination with a recombinant fusion protein encoding an astacinlike metalloprotease (MTP-1) secreted by host-stimulated Ancylostoma caninum third-stage infective larvae. *The Journal of parasitology* **89**, 853-5 (2003).
98. Pearson, M.S. et al. An enzymatically inactivated hemoglobinase from Necator americanus induces neutralizing antibodies against multiple hookworm species and protects dogs against heterologous hookworm infection. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **23**, 3007-19 (2009).
99. Ranjit, N. et al. A family of cathepsin B cysteine proteases expressed in the gut of the human hookworm, Necator americanus. *Molecular and biochemical parasitology* **160**, 90-9 (2008).

100. Loukas, A. et al. Vaccination of dogs with a recombinant cysteine protease from the intestine of canine hookworms diminishes the fecundity and growth of worms. *The Journal of infectious diseases* **189**, 1952-61 (2004).
101. Redmond, D.L. & Knox, D.P. Protection studies in sheep using affinity-purified and recombinant cysteine proteinases of adult Haemonchus contortus. *Vaccine* **22**, 4252-61 (2004).
102. Selzer, P.M. et al. Cysteine protease inhibitors as chemotherapy: lessons from a parasite target. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 11015-22 (1999).
103. Armstrong, P.B. Proteases and protease inhibitors: a balance of activities in host-pathogen interaction. *Immunobiology* **211**, 263-81 (2006).
104. Martin, J., Abubucker, S., Heizer, E., Taylor, C. & Mitreva, M. Nematode.net update 2011: addition of data sets and tools featuring next-generation sequencing data. *Nucleic acids research* **40**, D720 - D728 (2012).
105. Whitcomb, D.C. & Lowe, M.E. Human pancreatic digestive enzymes. *Digestive diseases and sciences* **52**, 1-17 (2007).
106. Milstone, A.M., Harrison, L.M., Bungiro, R.D., Kuzmic, P. & Cappello, M. A broad spectrum Kunitz type serine protease inhibitor secreted by the hookworm Ancylostoma ceylanicum. *The Journal of biological chemistry* **275**, 29391-9 (2000).
107. Hawdon, J.M., Datu, B. & Crowell, M. Molecular cloning of a novel multidomain Kunitz-type proteinase inhibitor from the hookworm Ancylostoma caninum. *The Journal of parasitology* **89**, 402-7 (2003).
108. Chu, D. et al. Molecular characterization of Ancylostoma ceylanicum Kunitz-type serine protease inhibitor: evidence for a role in hookworm-associated growth delay. *Infection and immunity* **72**, 2214-21 (2004).
109. Hartmann, S. & Lucius, R. Modulation of host immune responses by nematode cystatins. *International journal for parasitology* **33**, 1291-302 (2003).
110. Vray, B., Hartmann, S. & Hoebeke, J. Immunomodulatory properties of cystatins. *Cellular and molecular life sciences : CMLS* **59**, 1503-12 (2002).
111. Newlands, G.F., Skuce, P.J., Knox, D.P. & Smith, W.D. Cloning and expression of cystatin, a potent cysteine protease inhibitor from the gut of Haemonchus contortus. *Parasitology* **122**, 371-8 (2001).
112. Dainichi, T., Maekawa, Y., Ishii, K. & Himeno, K. Molecular cloning of a cystatin from parasitic intestinal nematode, Nippostrongylus brasiliensis. *The journal of medical investigation : JMI* **48**, 81-7 (2001).
113. Schonemeyer, A. et al. Modulation of human T cell responses and macrophage functions by onchocystatin, a secreted protein of the filarial nematode Onchocerca volvulus. *Journal of immunology* **167**, 3207-15 (2001).
114. Prufer, K. et al. FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC bioinformatics* **8**, 41 (2007).
115. Hu, M. et al. Structural and functional characterisation of the fork head transcription factor-encoding gene, Hc-daf-16, from the parasitic nematode Haemonchus contortus (Strongylida). *International journal for parasitology* **40**, 405-15 (2010).
116. Gao, X. et al. Identification of hookworm DAF-16/FOXO response elements and direct gene targets. *PloS one* **5**, e12289 (2010).

117. Huang, H. & Tindall, D.J. Dynamic FoxO transcription factors. *J Cell Sci* **120**, 2479-87 (2007).
118. Mihailovich, M., Militti, C., Gabaldon, T. & Gebauer, F. Eukaryotic cold shock domain proteins: highly versatile regulators of gene expression. *BioEssays : news and reviews in molecular, cellular and developmental biology* **32**, 109-18 (2010).
119. Tabach, Y. et al. Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. *PloS one* **2**, e807 (2007).
120. Koudritsky, M. & Domany, E. Positional distribution of human transcription factor binding sites. *Nucleic acids research* **36**, 6795-805 (2008).
121. Gerstein, M.B. et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**, 1775-87 (2010).
122. Chalmers, I.W. et al. Developmentally regulated expression, alternative splicing and distinct sub-groupings in members of the Schistosoma mansoni venom allergen-like (SmVAL) gene family. *BMC Genomics* **9**, 89 (2008).
123. Ranjit, N., Jones, M.K., Stenzel, D.J., Gasser, R.B. & Loukas, A. A survey of the intestinal transcriptomes of the hookworms, Necator americanus and Ancylostoma caninum, using tissues isolated by laser microdissection microscopy. *International journal for parasitology* **36**, 701-10 (2006).
124. Moreno, Y. et al. Proteomic analysis of excretory-secretory products of Heligmosomoides polygyrus assessed with next-generation sequencing transcriptomic information. *PLoS neglected tropical diseases* **5**, e1370 (2011).
125. Asojo, O.A. et al. X-ray structure of Na-ASP-2, a pathogenesis-related-1 protein from the nematode parasite, Necator americanus, and a vaccine antigen for human hookworm infection. *Journal of molecular biology* **346**, 801-14 (2005).
126. Bethony, J.M. et al. Randomized, placebo-controlled, double-blind trial of the Na-ASP-2 hookworm vaccine in unexposed adults. *Vaccine* **26**, 2408-17 (2008).
127. Milne, T.J., Abbenante, G., Tyndall, J.D., Halliday, J. & Lewis, R.J. Isolation and characterization of a cone snail protease with homology to CRISP proteins of the pathogenesis-related protein superfamily. *The Journal of biological chemistry* **278**, 31105-10 (2003).
128. Loukas, A. & Maizels, R.M. Helminth C-type lectins and host-parasite interactions. *Parasitology today* **16**, 333-9 (2000).
129. Hewitson, J.P., Grainger, J.R. & Maizels, R.M. Helminth immunoregulation: the role of parasite secreted proteins in modulating host immunity. *Molecular and biochemical parasitology* **167**, 1-11 (2009).
130. Loukas, A., Brown, A.P. & Pritchard, D.I. Na-ctl-2, a cDNA encoding a C-type lectin expressed exclusively in adult Necator americanus hookworms. *DNA sequence : the journal of DNA sequencing and mapping* **13**, 61-5 (2002).
131. Turner, D.G., Wildblood, L.A., Inglis, N.F. & Jones, D.G. Characterization of a galectin-like activity from the parasitic nematode, Haemonchus contortus, which modulates ovine eosinophil migration in vitro. *Veterinary immunology and immunopathology* **122**, 138-45 (2008).
132. Kasper, G. et al. A calreticulin-like molecule from the human hookworm Necator americanus interacts with C1q and the cytoplasmic signalling domains of some integrins. *Parasite immunology* **23**, 141-52 (2001).

133. Nagaraj, S.H., Gasser, R.B. & Ranganathan, S. Needles in the EST haystack: large-scale identification and analysis of excretory-secretory (ES) proteins in parasitic nematodes using expressed sequence tags (ESTs). *PLoS neglected tropical diseases* **2**, e301 (2008).
134. Saverwyns, H. et al. Identification and characterization of a novel specific secreted protein family for selected members of the subfamily Ostertagiinae (Nematoda). *Parasitology* **135**, 63-70 (2008).
135. Henkle-Duhrsen, K. & Kampkotter, A. Antioxidant enzyme families in parasitic nematodes. *Molecular and biochemical parasitology* **114**, 129-42 (2001).
136. Joachim, A., Lautscham, E., Christoffers, J. & Ruttkowski, B. Oesophagostomum dentatum: effect of glutathione S-transferase (GST) inhibitors on GST activity and larval development. *Experimental parasitology* **127**, 762-7 (2011).
137. Ouaissi, A., Ouaissi, M. & Sereno, D. Glutathione S-transferases and related proteins from pathogenic human parasites behave as immunomodulatory factors. *Immunology letters* **81**, 159-64 (2002).
138. Jariwala, A.R. et al. Potency testing for the experimental Na-GST-1 hookworm vaccine. *Expert review of vaccines* **9**, 1219-30 (2010).
139. Goud, G.N. et al. Expression of the Necator americanus hookworm larval antigen Na-ASP-2 in Pichia pastoris and purification of the recombinant protein for use in human clinical trials. *Vaccine* **23**, 4754-64 (2005).
140. Hobert, O. The neuronal genome of Caenorhabditis elegans. *WormBook*, 1-106 (2013).
141. Adelsberger, H., Lepier, A. & Dudel, J. Activation of rat recombinant alpha(1)beta(2)gamma(2S) GABA(A) receptor by the insecticide ivermectin. *Eur J Pharmacol* **394**, 163-70 (2000).
142. Chen, T. Nuclear receptor drug discovery. *Current opinion in chemical biology* **12**, 418-26 (2008).
143. Wang, Z. et al. Identification of the nuclear receptor DAF-12 as a therapeutic target in parasitic nematodes. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9138-43 (2009).
144. Antebi, A. Nuclear hormone receptors in C. elegans. *WormBook : the online review of C. elegans biology*, 1-13 (2006).
145. Sluder, A.E. & Maina, C.V. Nuclear receptors in nematodes: themes and variations. *Trends in genetics : TIG* **17**, 206-13 (2001).
146. Li, C. & Kim, K. Neuropeptides. *WormBook : the online review of C. elegans biology*, 1-36 (2008).
147. Marks, N.J. & Maule, A.G. Neuropeptides in helminths: occurrence and distribution. *Advances in experimental medicine and biology* **692**, 49-77 (2010).
148. Yeh, I., Hanekamp, T., Tsoka, S., Karp, P.D. & Altman, R.B. Computational analysis of Plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. *Genome research* **14**, 917-24 (2004).
149. Chen, G. et al. Discordant protein and mRNA expression in lung adenocarcinomas. *Molecular & cellular proteomics : MCP* **1**, 304-13 (2002).
150. Fu, X. et al. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* **10**, 1471-2164 (2009).
151. Guo, Y. et al. How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes. *Acta Biochim Biophys Sin* **40**, 426-36 (2008).

152. Orntoft, T.F., Thykjaer, T., Waldman, F.M., Wolf, H. & Celis, J.E. Genome-wide study of gene copy numbers, transcripts, and protein levels in pairs of non-invasive and invasive human transitional cell carcinomas. *Molecular & cellular proteomics : MCP* **1**, 37-45 (2002).
153. Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* **4**, 117 (2003).
154. Plieskatt, J.L. et al. Advances in vaccines against neglected tropical diseases: enhancing physical stability of a recombinant hookworm vaccine through biophysical and formulation studies. *Hum Vaccin Immunother* **8**, 765-76 (2012).
155. Takemoto, T. et al. Cloning and characterization of the Caenorhabditis elegans CeCRMP/DHP-1 and -2; common ancestors of CRMP and dihydropyrimidinase? *Gene* **261**, 259-67 (2000).
156. Kimura, J., Kudoh, T., Miki, Y. & Yoshida, K. Identification of dihydropyrimidinase-related protein 4 as a novel target of the p53 tumor suppressor in the apoptotic response to DNA damage. *International journal of cancer. Journal international du cancer* **128**, 1524-31 (2011).
157. Gounaris, K. Nucleotidase cascades are catalyzed by secreted proteins of the parasitic nematode Trichinella spiralis. *Infection and immunity* **70**, 4917-24 (2002).
158. Melnikov, A. et al. Clinical and environmental isolates of Burkholderia cepacia exhibit differential cytotoxicity towards macrophages and mast cells. *Molecular microbiology* **36**, 1481-93 (2000).
159. Giordani, R.B. et al. Trichomonas vaginalis nucleoside triphosphate diphosphohydrolase and ecto-5'-nucleotidase activities are inhibited by lycorine and candimine. *Parasitology international* **59**, 226-31 (2010).
160. Ghosh, I., Raghavan, N., FitzGerald, P.C. & Scott, A.L. Nucleoside diphosphate kinase from the parasitic nematode Brugia malayi. *Gene* **164**, 261-6 (1995).
161. Gounaris, K., Thomas, S., Najarro, P. & Selkirk, M.E. Secreted variant of nucleoside diphosphate kinase from the intracellular parasitic nematode Trichinella spiralis. *Infection and immunity* **69**, 3658-62 (2001).
162. Yatsuda, A.P., Krijgsveld, J., Cornelissen, A.W., Heck, A.J. & de Vries, E. Comprehensive analysis of the secreted proteins of the parasite Haemonchus contortus reveals extensive sequence variation and differential immune recognition. *The Journal of biological chemistry* **278**, 16941-51 (2003).
163. Craig, H., Wastling, J.M. & Knox, D.P. A preliminary proteomic survey of the in vitro excretory/secretory products of fourth-stage larval and adult Teladorsagia circumcincta. *Parasitology* **132**, 535-43 (2006).
164. Zimhony, O., Cox, J.S., Welch, J.T., Vilcheze, C. & Jacobs, W.R., Jr. Pyrazinamide inhibits the eukaryotic-like fatty acid synthetase I (FASI) of Mycobacterium tuberculosis. *Nature medicine* **6**, 1043-7 (2000).
165. Wakil, S.J. Fatty acid synthase, a proficient multifunctional enzyme. *Biochemistry* **28**, 4523-30 (1989).
166. Kennedy, J.A., Unger, S.A. & Horowitz, J.D. Inhibition of carnitine palmitoyltransferase-1 in rat heart and liver by perhexiline and amiodarone. *Biochemical pharmacology* **52**, 273-80 (1996).

167. Ceccarelli, S.M., Chomienne, O., Gubler, M. & Arduini, A. Carnitine palmitoyltransferase (CPT) modulators: a medicinal chemistry perspective on 35 years of research. *Journal of medicinal chemistry* **54**, 3109-52 (2011).
168. Mujtaba, Q. & Burrow, G.N. Treatment of hyperthyroidism in pregnancy with propylthiouracil and methimazole. *Obstetrics and gynecology* **46**, 282-6 (1975).
169. Czarnocka, B., Ruf, J., Ferrand, M., Carayon, P. & Lissitzky, S. Purification of the human thyroid peroxidase and its identification as the microsomal antigen involved in autoimmune thyroid diseases. *FEBS letters* **190**, 147-52 (1985).
170. Mones, R.J. An analysis of six patients with Parkinson's disease who have been unresponsive to L-dopa therapy. *Journal of neurology, neurosurgery, and psychiatry* **36**, 362-7 (1973).
171. Lovenberg, W., Weissbach, H. & Udenfriend, S. Aromatic L-amino acid decarboxylase. *The Journal of biological chemistry* **237**, 89-93 (1962).
172. Pearson, D.C., May, G.R., Fick, G.H. & Sutherland, L.R. Azathioprine and 6-mercaptopurine in Crohn disease. A meta-analysis. *Annals of internal medicine* **123**, 132-42 (1995).
173. Mason, M. et al. Azathioprine in rheumatoid arthritis. *British medical journal* **1**, 420-2 (1969).
174. Toth, E.A. & Yeates, T.O. The structure of adenylosuccinate lyase, an enzyme with dual activity in the de novo purine biosynthetic pathway. *Structure* **8**, 163-74 (2000).
175. Fitton, A. & Goa, K.L. Azelaic acid. A review of its pharmacological properties and therapeutic efficacy in acne and hyperpigmentary skin disorders. *Drugs* **41**, 780-98 (1991).
176. Wilson, J.D., Griffin, J.E. & Russell, D.W. Steroid 5 alpha-reductase 2 deficiency. *Endocrine reviews* **14**, 577-93 (1993).
177. Vigil, A., Davies, D.H. & Felgner, P.L. Defining the humoral immune response to infectious agents using high-density protein microarrays. *Future Microbiol* **5**, 241-51 (2010).