nature genetics
OPEN

Genome of the human hookworm *Necator americanus*

Yat T Tang[1,16], Xin Gao[1,16], Bruce A Rosa[1,16], 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 K 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]

The hookworm *Necator americanus* is the predominant soil-transmitted human parasite. Adult worms feed on blood in the small intestine, causing iron-deficiency anemia, malnutrition, growth and development stunting in children, and severe morbidity and mortality during pregnancy in women. We report sequencing and assembly of the *N. americanus* genome (244 Mb, 19,151 genes). Characterization of this first hookworm genome sequence identified genes orchestrating the hookworm's invasion of the human host, genes involved in blood feeding and development, and genes encoding proteins that represent new potential drug targets against hookworms. *N. americanus* has undergone a considerable and unique expansion of immunomodulator proteins, some of which we highlight as potential treatments against inflammatory diseases. We also used a protein microarray to demonstrate a postgenomic application of the hookworm genome sequence. This genome provides an invaluable resource to boost ongoing efforts toward fundamental and applied postgenomic research, including the development of new methods to control hookworm and human immunological diseases.

© 2014 Nature America, Inc. All rights reserved.

Soil-transmitted helminths (STHs), including *Ascaris*, *Trichuris* and hookworms, cause neglected tropical diseases affecting >1 billion people worldwide[1,2]. Hookworms alone infect approximately 700 million people, primarily in disadvantaged communities in tropical and subtropical regions, causing a disease burden of 1.5–22.1 million disability-adjusted life years[3]. *N. americanus* represents ~85% of all hookworm infections[4] and causes necatoriasis, characterized clinically by anemia, malnutrition in pregnant women, and an impairment of cognitive and/or physical development in children[5].

The life cycle of *N. americanus* commences with eggs being shed in the feces of infected people. Eggs embryonate in soil under favorable conditions, and then the first-stage larvae hatch, feed on environmental microbes and molt twice to become infective third-stage larvae (iL3). These larvae infect the human host by skin penetration, enter subcutaneous blood and lymph vessels, and travel via the circulation to the lungs. The iL3 break into the alveoli and migrate via the trachea to the oropharynx, after which they are swallowed and travel to the small intestine, where they develop to become dioecious adults. The adult worms (~1 cm long) attach to the mucosa, where they feed on blood (up to 30 µl per day per worm), and can survive in the human host for up to a decade. The pre-patent period of *N. americanus* is 4–8 weeks, and a female worm can produce up to 10,000 eggs per day.

New methods to control hookworm disease are urgently needed. Present therapy relies mainly on mass treatment with albendazole[6], but repeated and excessive use of this agent has the potential to lead to treatment failures[7] and drug resistance[8]. Recent indications of reduced cure rates in infected humans[9] imply an urgent need for new intervention strategies. Early attempts to use bioinformatic approaches for the discovery of immunogens were hampered by a lack of understanding of the molecular biology of *N. americanus* and other hookworms[4] and by the absence of genome and proteome sequences. A recent study[10] has shown that comparative genomics facilitates the characterization and prioritization of anthelmintic targets, which results in a higher hit rate than conventional approaches.

In addition to a need for anti-hookworm vaccines in countries with high rates of hookworm infections, hookworms and other helminths

[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, Queensland, Australia. [8]Queensland Institute of Medical Research, Brisbane, Queensland, 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. [16]These authors contributed equally to this work. Correspondence should be addressed to M.M. (mmitreva@genome.wustl.edu).

Received 10 June 2013; accepted 18 December 2013; published online 19 January 2014; doi:10.1038/ng.2875

Table 1 Summary of *N. americanus* genomic features

Estimated genome size (Mb)	244
Assembly statistics	
Total number of supercontigs (≥1 kb)	11,713
Total number of base pairs (bp) in supercontigs	244,009,025
Number of N50 supercontigs[a]	283
N50 supercontig length (bp)[a]	213,095
Number of N90 supercontigs[a]	1,336
N90 supercontig length (bp)[a]	29,214
GC content of whole genome	40.20%
Repetitive sequences	23.50%
Protein-coding loci	
Total number of protein-coding genes	19,151
Avg. gene locus footprint (bp)	4,289
Avg. number of exons per gene	6.4
Avg. exon size (bp)	125
Avg. intron size (bp)	642
Avg. intergenic space (bp)	6,631

[a]N50 and N90 respectively denote 50% and 90% of all nucleotides in the assembly. 50% of the genome is in 283 supercontigs and in supercontigs with a minimum length of 213 kb; 90% of the genome is in 1,336 supercontigs and in supercontigs with a minimum length of 29 kb.

© 2014 Nature America, Inc. All rights reserved.

npg

are being explored as treatments (probiotics) against immunological diseases in humans in many industrialized countries where hookworm infections are not endemic[11]. Recent studies[12–14] indicate that hookworms suppress the production of pro-inflammatory molecules and promote anti-inflammatory and wound-healing properties, suggesting a mechanism by which worms reside for long periods in humans and suppress autoimmune and allergic diseases. Indeed, hookworm recombinant proteins have been tested in clinical trials for non-infectious diseases[15].

We sequenced, assembled and characterized the *N. americanus* genome and compared it with those of other nematodes and the human host. Bioinformatic analyses of the protein-coding genes identified salient molecular groups, some of which may represent new intervention targets. The production and screening of a hookworm protein microarray revealed previously undescribed features of the immune response to the parasite and enabled a postgenomic exploration of the genome sequence. In the postgenomic analysis, we identified molecules that have low similarity to proteins in other species but are recognized by all infected individuals and therefore have high diagnostic potential.

RESULTS

Genome features

The nuclear genome of *N. americanus* (244 megabases (Mb)) was assembled, with 11.4% (1,336) of the supercontigs (≥1 kb) comprising 90% of the genome. The 244-Mb sequence was estimated to represent 92% of the *N. americanus* genome (**Table 1**, **Supplementary Figs. 1–3** and **Supplementary Note**). The GC content was 40.2%, the amino acid composition was comparable to that of other species (including five nematodes, the host and two outgroups; **Supplementary Table 1**) and the repeat content was 23.5%. In total, 669 repeat families were predicted and annotated (**Supplementary Table 2** and **Supplementary Note**). The protein-encoding genes predicted (n = 19,151) represent 33.7% of the genome at an average density of 78.5 genes per Mb and a GC content of 45.8%.

Compared to those of *Caenorhabditis elegans*, *N. americanus* exons were shorter and the introns were longer (**Fig. 1a**), but the average intron length and count for genes orthologous between the two species was not significantly different (P = 0.65 and 0.69, respectively; **Fig. 1a**,**b** and **Supplementary Note**). However, introns in *C. elegans* genes that were orthologous to *N. americanus* genes were significantly longer than introns in nonorthologous *C. elegans* genes ($P < 1 \times 10^{-15}$; **Fig. 1c**). This may indicate a diversity of function for these genes, as longer introns are thought to contain functional elements in addition to what might be regarded as 'normal' intron structure[16]. Furthermore, *N. americanus* iL3-overexpressed genes had longer introns than adult-overexpressed genes (**Fig. 1b**), which may indicate a greater diversity of regulation for these gene sets[16]. Positional bias was observed for intron length, which was comparable to *C. elegans* position-specific intron lengths for orthologous genes (**Fig. 1c** and **Supplementary Note**).

Most genes (82.6%) were confirmed using RNA sequencing (RNA-seq) data from the iL3 and adult stages of *N. americanus* (two biological replicates per stage), and 6.5% and 3.7% were overexpressed in these stages, respectively (**Supplementary Figs. 4** and **5**, and **Supplementary Table 3**). Alternative splicing was detected for 24.6% (4,712) of the genes, of which ~68.3% have orthologs in *C. elegans*. Among *N. americanus* genes with *C. elegans* orthologs, the alternatively spliced genes were more likely than other genes to belong to orthologous groups for which more than half of the *C. elegans* genes were also alternatively spliced (P = 0.037, binomial distribution test). As expected, genes associated with alternative splicing had a higher number of exons than those without ($P < 10^{-15}$ and 2×10^{-7} for *N. americanus* and *C. elegans*, respectively). A total of 3,223 *N. americanus* genes were predicted to be trans-spliced, of which 818 had conserved gene order and orientation with 373 *C. elegans* operons (**Fig. 1d**, **Supplementary Figs. 6** and **7**, **Supplementary Table 4** and **Supplementary Note**). The expression profiles of genes within operons were significantly more similar to one another than to those of random subsets of non-operon genes ($P < 0.0001$), supporting the idea that they are co-transcribed under similar regulatory control[17].

The *N. americanus* predicted secretome (classical secretion, 1,590 proteins; nonclassical secretion, 4,785 proteins) represented 33% of the deduced proteome. Functional annotation of predicted proteins on the basis of sequence comparisons identified 4,961 unique domains and 1,411 Gene Ontology terms for 57% and 44% of the *N. americanus* genes, respectively, and annotations were provided for 68% of the predicted *N. americanus* proteins (**Supplementary Table 5**).

Transcript expression in infective and parasitic stages

Hookworms spend a considerable amount of time as free-living larvae in the external environment before transitioning to parasitism. Differences in gene expression between these stages reflect this developmental progression (**Supplementary Table 3** and **Supplementary Fig. 5**). Of the 1,948 differentially expressed genes, 36% were significantly overexpressed (according to EdgeR, q = 0.05) in iL3, and 64% in adult. Compared to iL3-overexpressed genes, nearly twice as many of the adult-overexpressed genes were specific to *N. americanus* (58% compared to 32%, $P < 10^{-15}$), suggesting that species-specific genes are more likely to be related to parasitism rather than to the nonparasitic iL3 stage[18].

Among the iL3-overexpressed genes, eight molecular functions were over-represented ($P < 0.01$), including signal transduction, transmembrane receptor activity and anion transporter activity, reflecting the ability of iL3 to adapt to a complex environment and infect a suitable host (**Fig. 2a**, **Supplementary Table 6** and **Supplementary Note**). This finding is supported by the enrichment of genes encoding G protein–coupled receptor proteins among iL3-overexpressed genes ($P = 5.1 \times 10^{-8}$) but not among adult-overexpressed genes ($P = 4.1 \times 10^{-7}$) (**Supplementary Fig. 8**). Consistent with observations in other parasitic nematodes[19], serine/threonine protein kinase activity

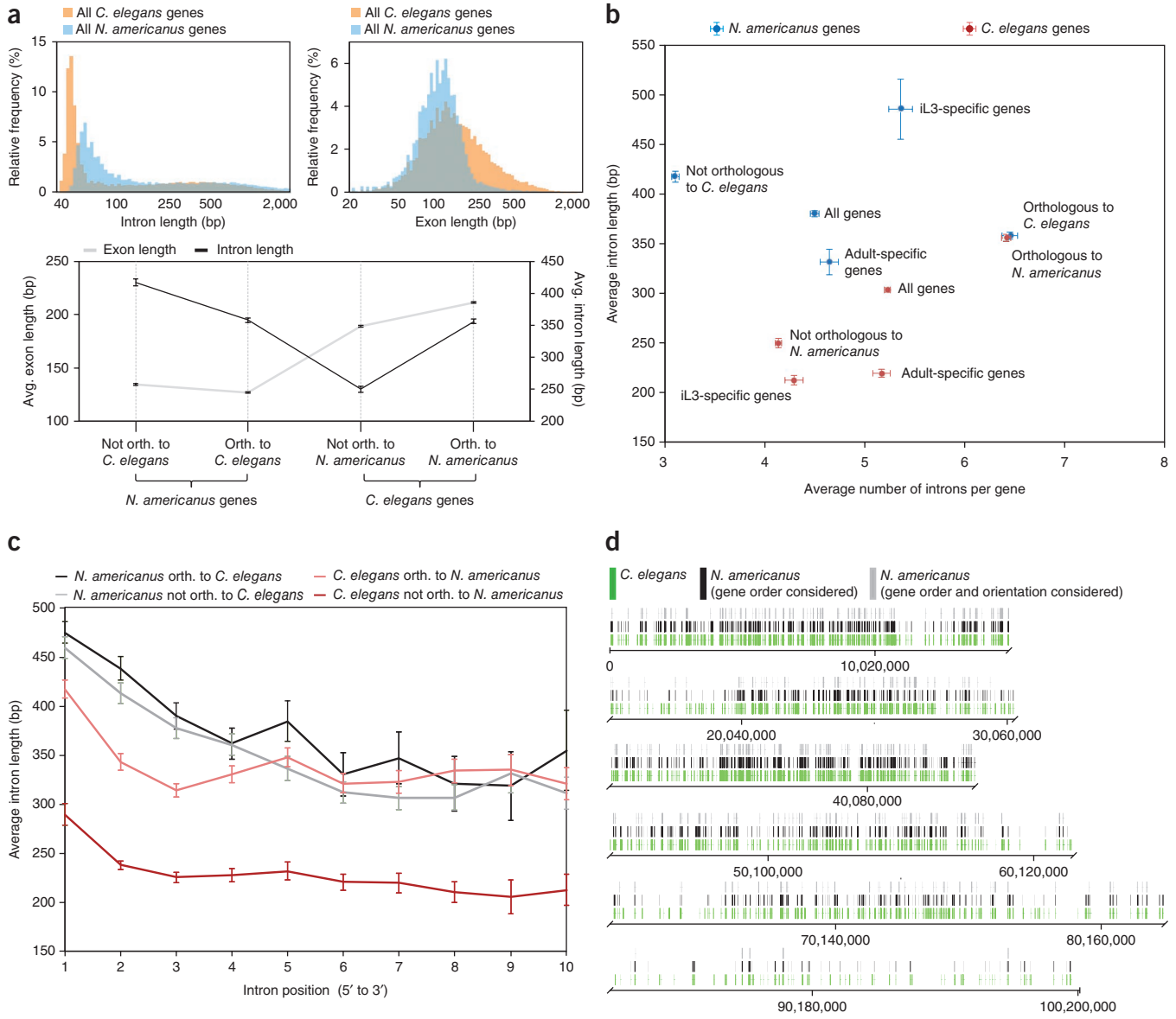

Figure 1 Organization of *N. americanus* gene features compared to *C. elegans*. (**a**) The average exon in *N. americanus* genes is significantly ($P < 1 \times 10^{-10}$) shorter and the average intron is significantly ($P < 1 \times 10^{-10}$) longer than in *C. elegans* genes. (**b**) Orthologous (orth.) genes have significantly ($P < 1 \times 10^{-10}$) more introns than nonorthologous genes in both species. (**c**) In orthologous genes from *C. elegans*, introns are longer at every intron position compared to nonorthologous genes. In **a**–**c**, error bars indicate s.e.m. (**d**) *N. americanus* genes that are in operons and conserved with *C. elegans* are shown on the *C. elegans* chromosomes.

was also enriched among iL3-overexpressed genes ($P = 0.008$). The complexity of transcription regulatory activities is likely to be high in iL3, as evidenced by the enrichment of genes annotated with "sequence-specific DNA binding transcription factor activity" (GO:0003700; $P = 1.7 \times 10^{-14}$) and genes with alternative splicing ($P < 2 \times 10^{-13}$), and by the fact that most (92.5%) of the differentially expressed transcription factors were iL3 overexpressed (**Supplementary Note**). This iL3-stage enrichment of transcription factor–related activity might indicate that transcription factors are poised for rapid gene expression after host invasion (that is, gene expression is not active but is likely to be primed, as observed in arrested stages of *C. elegans*[20]).

In contrast, in the adult stage, we detected overexpression of transcripts for a broad spectrum of enzymes including proteases, hydrolases and catalases (**Supplementary Table 6**). This reflects the nutritional adaptation of adult worms to a high-protein diet of blood[21] (**Fig. 2**, **Supplementary Fig. 9** and **Supplementary Note**). Proteins with a signal peptide (SP) for secretion had transcripts that were enriched among adult-overexpressed genes ($P < 10^{-15}$), whereas transmembrane domain–containing proteins ($P = 1.2 \times 10^{-8}$) had transcripts enriched among iL3-overexpressed genes. Proteases and protease inhibitors were enriched among SP-containing genes, and proteases contributed substantially to the predicted secretome (**Supplementary Table 6** and **Supplementary Note**), with 55% of all proteases (325 of 592) predicted to be secreted. Proteases, particularly *N. americanus*–specific proteases with no orthologs in *C. elegans*, were overexpressed more often in adult than in iL3 ($P < 10^{-15}$ for both comparisons; **Fig. 2b**,**c**, **Supplementary Note** and **Supplementary Table 7**). Serine-type endopeptidase inhibitor activity, required to protect the adult stage from the digestive and immunologically hostile environment in the host[22], was adult enriched ($P = 1.6 \times 10^{-4}$). The

© 2014 Nature America, Inc. All rights reserved.

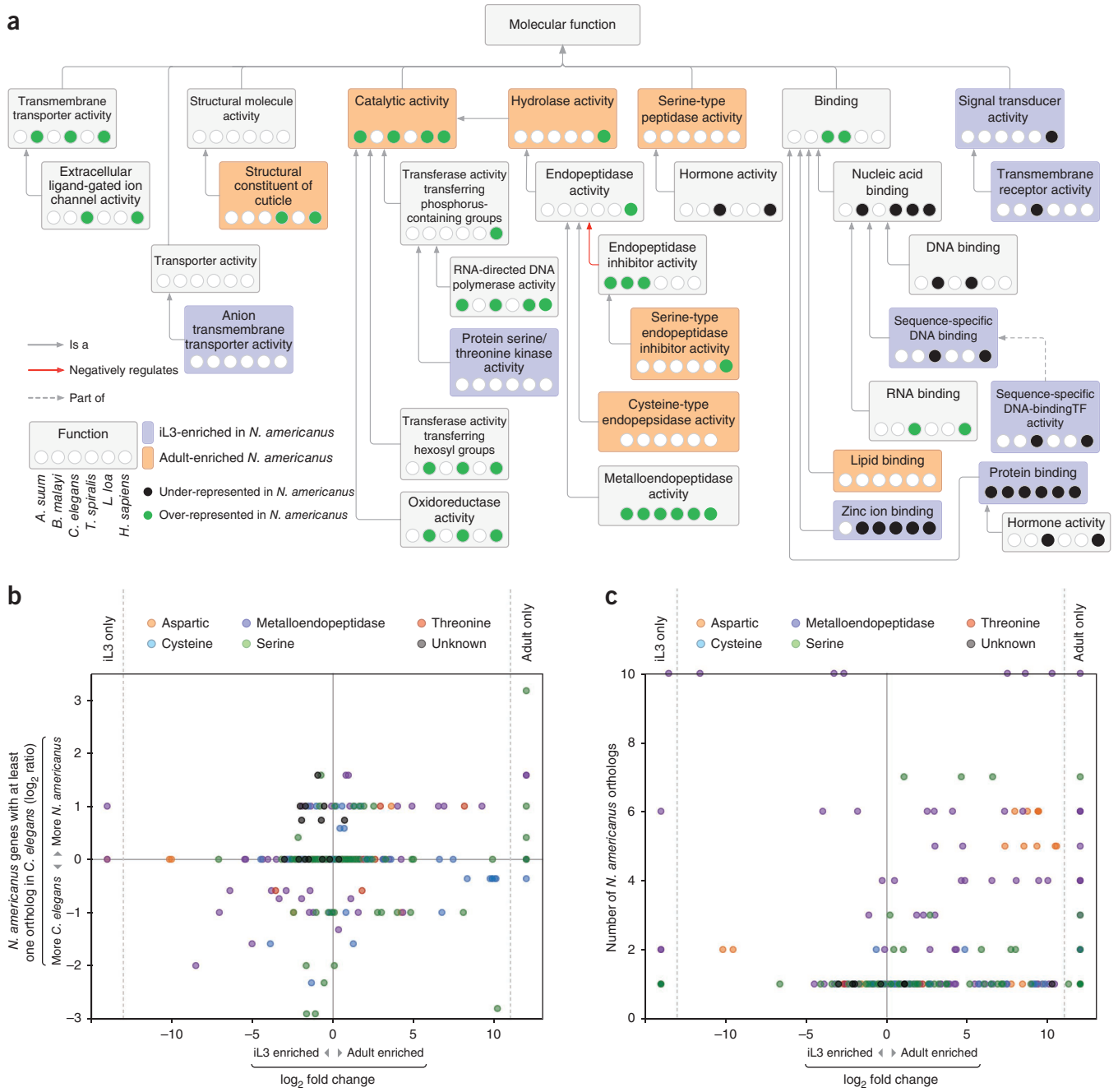

Figure 2 Molecular functions enriched among *N. americanus* genes, stage-enriched genes and the *N. americanus* degradome. (**a**) 'Molecular function' gene ontology terms enriched in specific life-cycle stages and in *N. americanus* compared to other species. Included are (i) categories enriched in the iL3 or adult life cycle stages in *N. americanus*, (ii) categories significantly ($P \leq 1 \times 10^{-5}$) over-represented or depleted in *N. americanus* compared to at least two of the comparison species, and (iii) second-order root nodes. TF, transcription factor. (**b**,**c**) Expression profiling of *N. americanus* proteases with *C. elegans* orthologs (**b**) or with no *C. elegans* orthologs (**c**).

adult enrichment of genes encoding structural constituents of the cuticle ($P = 1.7 \times 10^{-5}$) also relates to protecting the parasite from the host[23].

Blood feeding in adult hookworms is facilitated by an anticoagulation process and degradation of blood proteins by proteases. Known hookworm anticoagulants[24] are dominated by single-domain serine protease inhibitors (SPIs). We annotated 87 SPIs in *N. americanus*, accounting for 8 of 17 protease inhibitor clans. Given that serine proteases in humans are involved in diverse physiological functions, including blood coagulation and immunomodulation, the diversity of SPIs in *N. americanus* is probably crucial not only for anticoagulation during blood feeding but also for long-term survival in the host. Specifically, SPIs are likely to protect adult worms from enzymes in the small intestine, where serine proteases, including trypsin, chymotrypsin and elastase, are prominent[25], thus mediating hookworm-associated growth delay[22]. SPIs were enriched among the adult-overexpressed genes ($P = 3.9 \times 10^{-8}$), but not among the iL3-overexpressed genes ($P = 0.35$). Most of the SPIs characterized in hookworms were Kunitz-type molecules (**Supplementary Note**), but our findings suggest that multiple types of SPIs are produced by adult *N. americanus* in the human host. A mass spectrometry–based

© 2014 Nature America, Inc. All rights reserved.

proteomics analysis was performed using whole adult *N. americanus* worms (Online Methods), and the proteins detected (**Supplementary Table 7** and **Supplementary Fig. 10**) were also enriched for proteases ($P = 4.9 \times 10^{-7}$) and SPIs ($P = 1.8 \times 10^{-4}$), as well as proteins with SPs ($P = 4.7 \times 10^{-11}$) and proteins representing a wide range of Gene Ontology terms, many related to proteolysis (**Supplementary Table 6** and **Supplementary Note**).

Pathogenesis and immunobiology of hookworm disease

N. americanus causes chronic disease and does not usually induce sterile immunity in the host. Adult hookworms are able to live in the host for several years because of their ability to modulate and evade host immune defenses[13] with their excretory-secretory products, which sustain development and create a site of immune privilege[26]. By comparing the *N. americanus* genome with genomes from other nematodes, its host and distant species, we identified molecules that facilitate parasitism. Sixty percent of *N. americanus* genes had an ortholog in the other species studied (**Supplementary Table 8**, **Supplementary Fig. 11** and **Supplementary Note**). Comparative analysis identified metalloendopeptidases as the most prominent *N. americanus* proteases (**Fig. 2a**); these proteases are probably associated with the cleavage of eotaxin and inhibition of eosinophil recruitment[27], in addition to tissue penetration[28] and hemoglobinolysis[29]. *N. americanus* is the only blood-feeding nematode included in the comparison, and the hierarchical structure for enriched molecular functions (**Fig. 2a**) revealed shared and unique patterns and subsequent functional relationships.

SCP/Tpx-1/Ag5/PR-1/Sc7 (SCP/TAPS; InterPro IPR014044; **Supplementary Table 5**) is a protein family inferred to be involved in host-parasite interactions (**Supplementary Note**). There were 137 SCP/TAPS proteins in *N. americanus*, representing a fourfold expansion of this protein family compared to other nematodes. More than half (69 of 137) of the *N. americanus* SCP/TAPS proteins were adult overexpressed ($P < 10^{-15}$; **Fig. 3a**), and only 6 of the 137 *N. americanus* SCP/TAPS proteins had orthologs in *C. elegans* (according to Markov clustering (MCL); see Online Methods). The presence of a limited repertoire of orthologs in *C. elegans* suggests that nematode SCP/TAPS proteins may have originated before parasitism. Primary sequence similarity classified SCP/TAPS proteins into multiple groups (**Fig. 3b**,**c** and **Supplementary Fig. 12**), only some of which contained *C. elegans* members, suggesting independent expansion of SCP/TAPS proteins after parasite speciation. The large expansion of SCP/TAPS proteins in *N. americanus* suggests multiple, possibly distinct roles in host-parasite interactions. SCP/TAPS proteins have been studied extensively as hookworm drug or vaccine candidates[30] and as therapeutics for human inflammatory diseases[15] or stroke[31] (**Supplementary Note**). The 96 *N. americanus*–specific SCP/TAPS identified here might serve as candidates for selective drug or vaccine targets[32] (**Supplementary Table 5**).

We identified a total of 336 *N. americanus* genes that are orthologous to previously predicted genes encoding immunogenic/immunomodulatory proteins in *Ascaris suum*[24], along with three homologs of genes encoding transforming growth factor-β (TGF-β), an important protein in modulation of inflammation and the evolution of nematode parasitism[33] (**Supplementary Table 5**). Additional genes in *N. americanus* encoding proteins inferred to be involved in host-parasite immunomodulatory interactions include macrophage migration inhibitory factor (MIF), neutrophil inhibitory factor (NIF), hookworm platelet inhibitor (HPI), galectins, C-type lectins (C-TL), peroxiredoxins (PRX) and glutathione S-transferases (GST), among others (**Supplementary Note**).

Prospects for new interventions

Historically, anthelmintic drugs have been discovered using *in vivo* and *in vitro* compound screens[34]. Recent comparative 'omics' studies (accompanied by experimental screening) in multiple nematode species[10] have shown that genomic and transcriptomic data can be used to prioritize targets and raise the hit rate compared with conventional approaches. Hence, the availability of the *N. americanus* genome is expected to enable comparative genomic and chemogenomic studies for the prediction and prioritization of therapeutic targets. As more than half (53%) of all current drug targets[35] consist of rhodopsin-like G protein–coupled receptors (GPCRs), nuclear receptors (NRs), ligand-gated ion channels (LGICs), kinases and voltage-gated ion channels (VGICs), we investigated these protein groups in the *N. americanus* genome to identify potential therapeutic targets (**Supplementary Table 9** and **Supplementary Note**).

GPCRs are attractive drug targets owing to their importance in signal transduction[35]. We identified 272 GPCR genes in *N. americanus*, whereas there are nearly 1,700 GPCR genes in *C. elegans*. Although GPCRs are challenging to characterize at the primary sequence level and the *N. americanus* genome is in a draft state, there may be a biological explanation for this difference in the number of GPCRs identified, including frequent amplifications of several subfamilies of GPCRs in *C. elegans* relative to the closely related *Caenorhabditis briggsae*[36]. Three of the five GRAFS families of GPCRs (glutamate, rhodopsin and frizzled, but not adhesion or secretin) were found in *N. americanus*. The putative GPCRs were enriched for iL3 overexpression (30 genes; $P = 5.1 \times 10^{-8}$), with only one gene being adult overexpressed ($P = 4.1 \times 10^{-7}$ for under-representation). *N. americanus* encodes members of both major ion-channel categories (LGICs and VGICs); 224 LGICs belonging to two of the three subfamilies of LGIC (Cys-loop family and glutamate-activated cation channels) were identified, compared with 159 LGIC-encoding genes in *C. elegans*[37]. Genes encoding nicotinic acetylcholine receptor subunits (nAChR) of Cys-loop family members were also found. Nematodes have a much larger number of nAChR-α subunits than examined vertebrates (17 nAChR-encoding genes in mammals and birds, compared with 29 nAChR subunits in *C. elegans*[38]), and several anthelmintics such as levamisole[39] and monepantel[40] have been developed to exploit these differences. Ivermectin[41] targets a subunit of glutamate-gated chloride channels that are present in *N. americanus* (eight genes; InterPro IPR015680); three of these genes clustered with six *C. elegans* glutamate-gated chloride channel genes (*avr-14*, *avr-15* and *glc-1* to *glc-4*; ref. 42). The lack of a clear ortholog of the ivermectin-sensitive genes within the *N. americanus* genome, and the underlying sequence diversity at a position correlated with direct activation by ivermectin, may explain the relative ivermectin insensitivity of *N. americanus*[43] (**Supplementary Note** and **Supplementary Fig. 13**) compared to other nematodes[44].

VGICs include sodium, potassium and calcium channels and are anthelmintic targets (for example, emodepside inhibits SLO-1 in *C. elegans*[45] and parasitic nematodes such as *A. suum*[46]). *N. americanus* encodes 48 VGICs (fewer than *C. elegans*), including members from the major families such as 6-transmembrane (6TM) potassium channels, voltage-gated calcium channels and voltage-gated chloride channels (**Supplementary Note**). As in other nematodes[47], voltage-gated sodium channels were not present in *N. americanus*.

Protein kinases are involved in numerous signal transduction pathways that regulate biological processes, and they have been a

© 2014 Nature America, Inc. All rights reserved.

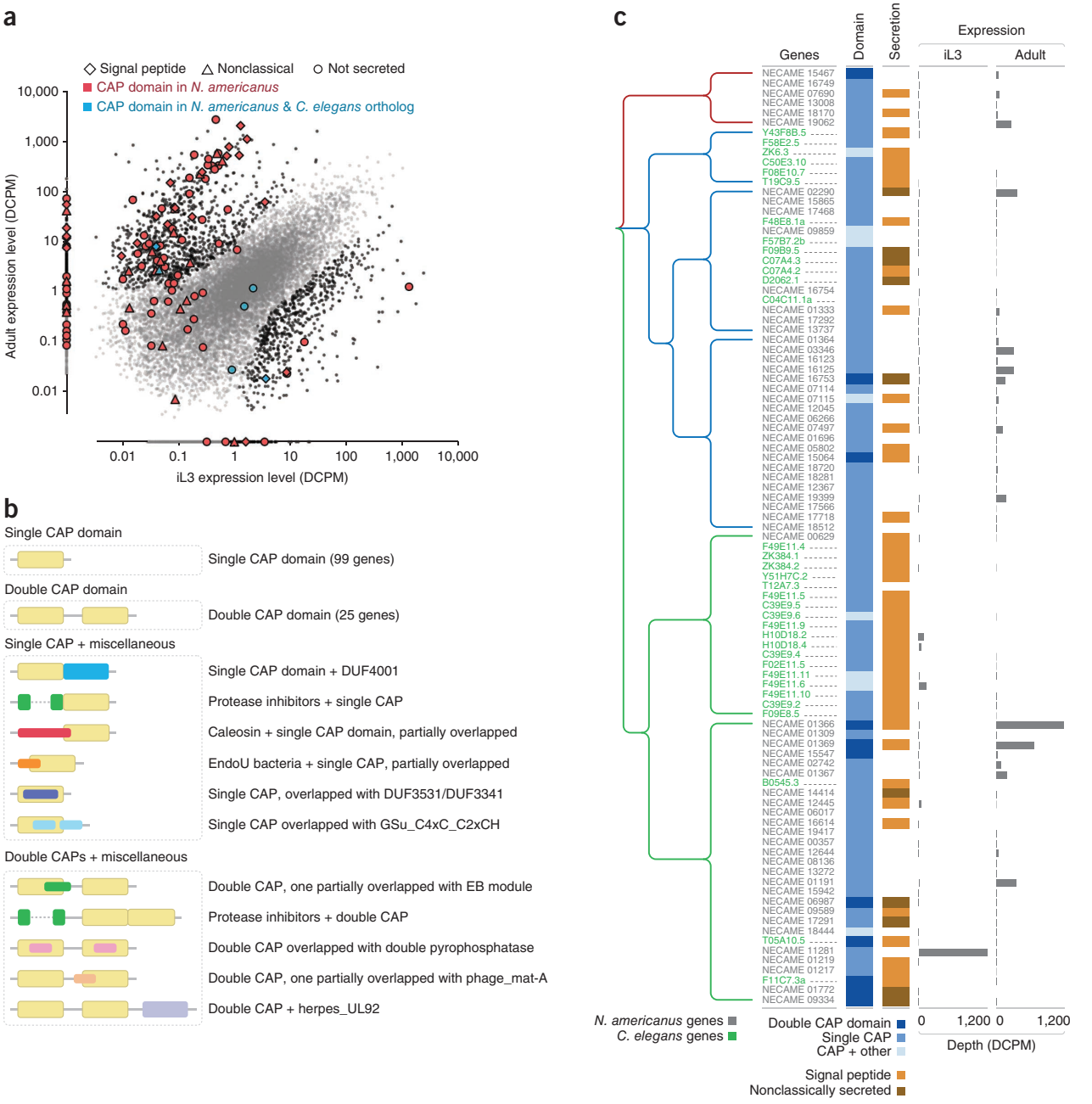

Figure 3 SCP/TAPS gene family expansion in *N. americanus*. (**a**) SCPs/TAPS are enriched in the adult stage of *N. americanus*. (**b**) Schematic representation of gene structure from SCP/TAPS family members. All SCP/TAPS proteins are grouped according to their number of CAP domains and regions outside the CAP domains: single CAP domain, double CAP domain, single CAP plus miscellaneous, and double CAP plus miscellaneous. (**c**) Neighbor-joining clustering of the all *C. elegans* and ungapped *N. americanus* SCP/TAPS genes on the basis of their full-length primary sequence similarity of the CAP domain. Data on domain representation, secretion type and stage of expression are included.

major focus for drug discovery[48,49]. Of the 274 *N. americanus* genes encoding kinases, 15 and 12 were overexpressed in iL3 and adults, respectively. Gene expression, tissue expression, conservation among nematodes and dissimilarity to human orthologs were used for prioritization[10] of candidate targets (**Supplementary Table 10**). To evaluate current drugs and inhibitors that target homologous kinases, we also prioritized compounds from a publicly available database (Online Methods). The highest-scoring compound was a tyrosine kinase inhibitor approved for treating chronic myelogenous leukemia[9]. A total of 233 other compounds had the second-highest score of 5 (**Supplementary Table 11**), indicating that these existing drugs might be repurposed for treating neglected tropical diseases, thus minimizing development time and cost[50].

Chokepoints in metabolic pathways[51] were analyzed and prioritized to identify further drug targets. *N. americanus* encodes at least 3,976 protein-coding genes associated with 3,265 KEGG orthology terms (**Supplementary Table 7**), 938 (24%) of which are involved in metabolic pathways (**Supplementary Fig. 14**), representing 32 potentially complete modules. A total of 34% of the metabolic pathway genes were classified as chokepoints (**Supplementary Table 12**), of which 120 were conserved among nematodes and non-nematode species used in the comparative analysis. Chokepoint prioritization,

© 2014 Nature America, Inc. All rights reserved.

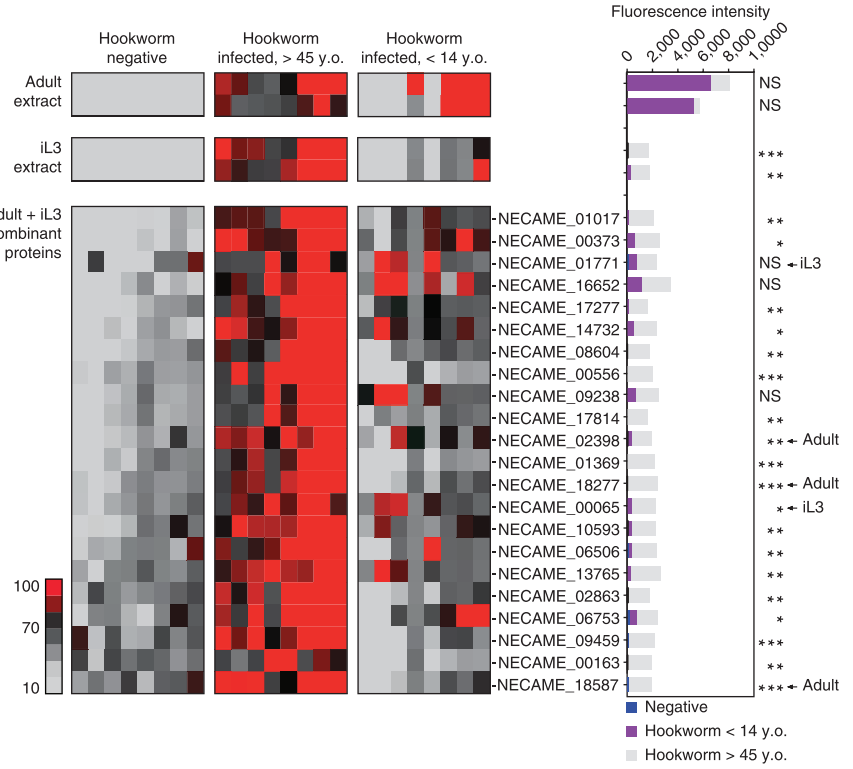

Figure 4 Serum responses to *N. americanus* antigens vary with age and infection intensity. The heat map shows the immunoreactivity of 22 antigens to the IgG antibodies from groups of uninfected individuals, infected children <14 years old and infected adults >45 years old ($n = 8$ in each group). Duplicate crude somatic extracts from iL3 and adult stages were included as control naive antigens. Every other row represents an individual recombinant *in vitro* translation product. The bar chart shows the mean immunoreactivity of the three groups for each antigen, measured by mean fluorescence intensity. 'iL3' and 'Adult' labels denote stage-specific expression of indicated antigens, measured by RNA-seq data. Significant differences in antibody responses between human adults and children were detected with Student's *t*-test. $*P < 0.05$; $**P < 0.01$; $***P < 0.001$; NS, no significant difference.

along with requirements for a chokepoint to be an expression bottleneck in *N. americanus* and for lethality upon RNAi knockdown of the orthologous gene in *C. elegans*, prioritized eight enzymes encoded by ten distinct genes (**Supplementary Tables 12–14** and **Supplementary Note**). Among the prioritized chokepoints is adenylosuccinate lyase (ASL; EC 4.3.2.2) (**Supplementary Figs. 15–17**), an enzyme involved in the purine metabolism pathway (KEGG pathway ko00230) and a chokepoint in the adenine ribonucleotide biosynthesis module (KEGG pathway M00049). To identify chokepoint inhibitors for repurposing, we assessed compounds from publicly available databases (449 target-compound pairs) using the same method as for kinase inhibitors. The highest-ranked candidates include compounds such as azathioprine (DrugBank DB00993), a prodrug that is converted into mercaptopurine (DrugBank DB01033) to inhibit purine metabolism and DNA synthesis (**Supplementary Fig. 18**, **Supplementary Table 14** and **Supplementary Note**).

Postgenomic exploration using the *N. americanus* immunome

The *N. americanus* genome enables development of postgenomic tools to investigate the immunobiology of human hookworm disease and accelerate antigen discovery for the development of vaccines and diagnostics. We developed a protein microarray containing 564 *N. americanus* recombinant proteins inferred from the genome (**Supplementary Table 15** and **Supplementary Note**). The microarray was probed with sera from individuals aged 4–66 years who were residents in an *N. americanus*–endemic area of northeastern Minas Gerais state in Brazil. This pilot study based on 200 individuals from the youngest (<14 years of age) and the oldest (>45 years of age) age strata identified 22 antigens that were significant ($P \leq 0.05$) targets of anti-hookworm immune responses (**Fig. 4**).

Older individuals showed stronger immunoglobin G (IgG) responses to a larger number of secreted antigens, but these antibodies seem to have no role in killing the parasite or protecting against heavy infection. Hence, unlike other STHs of humans, protective immunity to *N. americanus* does not seem to develop in most individuals during adolescence. This is consistent with observations that, in *Necator*-endemic areas, older people often harbor the heaviest-intensity infections[1,52,53]. Younger individuals showed IgG responses against fewer antigens, usually with lower intensity. Thus, although antibodies are a key feature of the immune response to *N. americanus* and increase with host age, they do not protect individuals from infection over time.

There are probably multiple factors contributing to the absence of overall protective immunity to hookworm infection, in contrast to the age-acquired protective immunity observed with other STH infections. Detailed kinetic studies of the IgG subclasses and IgE responses to hookworm antigens represented on our protein microarray will be required to better understand the roles of these antibodies in the acquisition of immunity against hookworm[13]. The protein microarray can be probed with sera from individuals with different genetic backgrounds and different histories of exposure to hookworm[54], as well as from animals rendered immunologically resistant to hookworm infection by vaccination with irradiated iL3 (ref. 55), thereby facilitating efforts to develop an efficacious vaccine against hookworm disease. Furthermore, secreted proteins that are recognized by most or all the infected individuals, and have weak or no homology to other nematode species, represent antigens that might form the basis of sensitive and specific serodiagnostic tests (**Supplementary Note**; for example, **Supplementary Fig. 19**).

DISCUSSION

N. americanus is responsible for causing more disease worldwide than any other STH. The characterization of the first genome of a human hookworm is expected to facilitate future fundamental explorations of the epidemiology and evolutionary biology of hookworms as well as efforts toward the development of therapeutics to combat hookworm disease. As *N. americanus* is the first hookworm whose genome has been sequenced, the data presented here provide a first insight into blood-feeding nematodes of major importance for human and animal health.

Our postgenomic exploration of inferred proteomic information highlights the utility of the draft genome sequence for understanding the immunobiology of human hookworm disease and accelerating

© 2014 Nature America, Inc. All rights reserved.

the development of vaccines and diagnostics. It is also pertinent to note that hookworms are garnering interest for their therapeutic properties against a range of noninfectious inflammatory diseases of humans. The genome sequence therefore represents a veritable pharmacopoeia—indeed, recombinant hookworm molecules have already undergone clinical trials for stroke and deep-vein thrombosis[15]. Thus, the *N. americanus* genome sequence will have broad implications. It provides many opportunities to establish postgenomic methods in the quest to develop improved interventions against this ancient and neglected parasite, as well as inflammatory diseases that are reaching epidemic proportions in industrialized societies.

URLs. NCBI SRA, http://www.ncbi.nlm.nih.gov/sra; RepeatModeler, http://www.repeatmasker.org/RepeatModeler.html; RNAmmer, http://www.cbs.dtu.dk/services/RNAmmer/; Rfam database, http://www.sanger.ac.uk/resources/databases/rfam.html; RepeatMasker, http://repeatmasker.org/; Fgenesh, www.softberry.com/; BER, http://ber.sourceforge.net/; Seqclean, http://compbio.dfci.harvard.edu/tgi/software/; Refcov, http://gmt.genome.wustl.edu/genome-shipit/gmt-refcov/current/; PyMOL, www.pymol.org/; KEGG transcription factor database, http://www.genome.jp/kegg-bin/get_htext?ko03000.keg; Jaspar database, http://jaspar.genereg.net/; Patser, http://stormo.wustl.edu/resources.html; Kinomer, http://www.compbio.dundee.ac.uk/kinomer; SignalP, www.cbs.dtu.dk/services/SignalP/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The whole-genome sequence of *N. americanus* has been deposited in DDBJ/EMBL/GenBank under the project accession ANCG00000000. The version described in this paper is the first version, ANCG01000000. All short-read data have been deposited in the Short Read Archive under the following accessions: SRR036799, SRR036800, SRR036802, SRR036804, SRR036811, SRR341459, SRR341460, SRR609850, SRR609895, SRR609951, SRR610281, SRR610282, SRR611341, SRR611350. RNA-seq profiles have been deposited in Nematode.net and a browsable genome is also available at Nematode.net and WormBase.

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

ACKNOWLEDGMENTS

We thank the faculty and staff of the Genome Institute at Washington University and the Protein Microarray Laboratory at the University of California–Irvine (U54AI065359) who contributed to this study. The genome sequencing and annotation work was funded by US National Institutes of Health (NIH)–National Human Genome Research Institute grant U54HG003079 to R.K.W. Comparative genome analysis was funded by grants NIH–National Institute of Allergy and Infectious Diseases AI081803 and NIH–National Institute of General Medical Sciences GM097435 to M.M. Funds from the Australian Research Council and Australia's National Health and Medical Research Council to R.B.G. are gratefully acknowledged. P.W.S. is an investigator with the Howard Hughes Medical Institute. We thank the faculty and staff of The Genome Institute at Washington University who contributed to this study.

AUTHOR CONTRIBUTIONS

Y.T.T., X.G. and B.A.R. contributed equally to this work. M.M., R.B.G., P.W.S., R.K.W. and S.R. conceived and planned the project. M.M. led the project, analysis and manuscript preparation. B.Z., P.J.H., J.M.H., P.L.F., J.B. and E.M.R. provided material. K.H.-P., X.Z., V.B.-P., P.M., W.C.W., J. Martin and S.A. produced sequence data and constructed, annotated and submitted the assembly. M.M., Y.T.T., X.G., B.A.R., R.T., Q.W., S.A., J. Martin, E.H., A.L., S.T.G., P.L.F., J. Mulvenna, J.S. and A.D. performed genome-based comparative studies, differential transcription, host-parasite interaction analysis, and proteomics and protein-array analysis. M.M., R.B.G., A.L. and J.M.H. drafted, edited and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/.

1. Bethony, J. *et al.* Soil-transmitted helminth infections: ascariasis, trichuriasis, and hookworm. *Lancet* **367**, 1521–1532 (2006).
2. Schneider, B. *et al.* A history of hookworm vaccine development. *Hum. Vaccin.* **7**, 1234–1244 (2011).
3. Hotez, P.J., Bethony, J.M., Diemert, D.J., Pearson, M. & Loukas, A. Developing vaccines to combat hookworm infection and intestinal schistosomiasis. *Nat. Rev. Microbiol.* **8**, 814–826 (2010).
4. Loukas, A. *et al.* Vaccinomics for the major blood feeding helminths of humans. *OMICS* **15**, 567–577 (2011).
5. Diemert, D.J., Bethony, J.M. & Hotez, P.J. Hookworm vaccines. *Clin. Infect. Dis.* **46**, 282–288 (2008).
6. Steinmann, P. *et al.* Efficacy of single-dose and triple-dose albendazole and mebendazole against soil-transmitted helminths and *Taenia* spp.: a randomized controlled trial. *PLoS ONE* **6**, e25003 (2011).
7. Keiser, J. & Utzinger, J. Efficacy of current drugs against soil-transmitted helminth infections: systematic review and meta-analysis. *J. Am. Med. Assoc.* **299**, 1937–1948 (2008).
8. Jia, T.W., Melville, S., Utzinger, J., King, C.H. & Zhou, X.N. Soil-transmitted helminth reinfection after drug treatment: a systematic review and meta-analysis. *PLoS Negl. Trop. Dis.* **6**, e1621 (2012).
9. Soukhathammavong, P.A. *et al.* Low efficacy of single-dose albendazole and mebendazole against hookworm and effect on concomitant helminth infection in Lao PDR. *PLoS Negl. Trop. Dis.* **6**, e1417 (2012).
10. Taylor, C.M. *et al.* Using existing drugs as leads for broad spectrum anthelmintics targeting protein kinases. *PLoS Pathog.* **9**, e1003149 (2013).
11. Elliott, D.E. & Weinstock, J.V. Helminth-host immunological interactions: prevention and control of immune-mediated diseases. *Ann. NY Acad. Sci.* **1247**, 83–96 (2012).
12. Daveson, A.J. *et al.* Effect of hookworm infection on wheat challenge in celiac disease–a randomised double-blinded placebo controlled trial. *PLoS ONE* **6**, e17366 (2011).
13. McSorley, H.J. & Loukas, A. The immunology of human hookworm infections. *Parasite Immunol.* **32**, 549–559 (2010).
14. Ferreira, I. *et al.* Hookworm excretory/secretory products induce interleukin-4 (IL-4)$^+$ IL-10$^+$ CD4$^+$ T cell responses and suppress pathology in a mouse model of colitis. *Infect. Immun.* **81**, 2104–2111 (2013).
15. Navarro, S., Ferreira, I. & Loukas, A. The hookworm pharmacopoeia for inflammatory diseases. *Int. J. Parasitol.* **43**, 225–231 (2013).
16. Bradnam, K.R. & Korf, I. Longer first introns are a general property of eukaryotic gene structure. *PLoS ONE* **3**, e3093 (2008).
17. Lercher, M.J., Blumenthal, T. & Hurst, L.D. Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. *Genome Res.* **13**, 238–243 (2003).
18. Wang, Z. *et al.* Characterizing *Ancylostoma caninum* transcriptome and exploring nematode parasitic adaptation. *BMC Genomics* **11**, 307 (2010).
19. Campbell, B.E., Hofmann, A., McCluskey, A. & Gasser, R.B. Serine/threonine phosphatases in socioeconomically important parasitic nematodes—prospects as novel drug targets? *Biotechnol. Adv.* **29**, 28–39 (2011).
20. Baugh, L.R., Demodena, J. & Sternberg, P.W. RNA Pol II accumulates at promoters of growth genes during developmental arrest. *Science* **324**, 92–94 (2009).
21. Williamson, A.L., Brindley, P.J., Knox, D.P., Hotez, P.J. & Loukas, A. Digestive proteases of blood-feeding nematodes. *Trends Parasitol.* **19**, 417–423 (2003).
22. Chu, D. *et al.* Molecular characterization of *Ancylostoma ceylanicum* Kunitz-type serine protease inhibitor: evidence for a role in hookworm-associated growth delay. *Infect. Immun.* **72**, 2214–2221 (2004).
23. Page, A.P. & Winter, A.D. Enzymes involved in the biogenesis of the nematode cuticle. *Adv. Parasitol.* **53**, 85–148 (2003).
24. Jex, A.R. *et al. Ascaris suum* draft genome. *Nature* **479**, 529–533 (2011).
25. Whitcomb, D.C. & Lowe, M.E. Human pancreatic digestive enzymes. *Dig. Dis. Sci.* **52**, 1–17 (2007).
26. Maizels, R.M. & Yazdanbakhsh, M. Immune regulation by helminth parasites: cellular and molecular mechanisms. *Nat. Rev. Immunol.* **3**, 733–744 (2003).
27. Culley, F.J. *et al.* Eotaxin is specifically cleaved by hookworm metalloproteases preventing its action *in vitro* and *in vivo*. *J. Immunol.* **165**, 6447–6453 (2000).
28. Kumar, S. & Pritchard, D.I. Secretion of metalloproteases by living infective larvae of *Necator americanus*. *J. Parasitol.* **78**, 917–919 (1992).

© 2014 Nature America, Inc. All rights reserved.

29. Ranjit, N. *et al.* Proteolytic degradation of hemoglobin in the intestine of the human hookworm *Necator americanus*. *J. Infect. Dis.* **199**, 904–912 (2009).
30. 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**, 4754–4764 (2005).
31. Krams, M. *et al.* Acute Stroke Therapy by Inhibition of Neutrophils (ASTIN): an adaptive dose-response study of UK-279,276 in acute ischemic stroke. *Stroke* **34**, 2543–2548 (2003).
32. Cantacessi, C. & Gasser, R.B. SCP/TAPS proteins in helminths—where to from now? *Mol. Cell. Probes* **26**, 54–59 (2012).
33. Viney, M.E., Thompson, F.J. & Crook, M. TGF-β and the evolution of nematode parasitism. *Int. J. Parasitol.* **35**, 1473–1475 (2005).
34. Kotze, A.C. Target-based and whole-worm screening approaches to anthelmintic discovery. *Vet. Parasitol.* **186**, 118–123 (2012).
35. Overington, J.P., Al-Lazikani, B. & Hopkins, A.L. How many drug targets are there? *Nat. Rev. Drug Discov.* **5**, 993–996 (2006).
36. Robertson, H.M. & Thomas, J.H. The putative chemoreceptor families of *C. elegans*. *WormBook* **2006**, 1–12 (2006).
37. Littleton, J.T. & Ganetzky, B. Ion channels and synaptic organization: analysis of the *Drosophila* genome. *Neuron* **26**, 35–43 (2000).
38. Jones, A.K., Davis, P., Hodgkin, J. & Sattelle, D.B. The nicotinic acetylcholine receptor gene family of the nematode *Caenorhabditis elegans*: an update on nomenclature. *Invert. Neurosci.* **7**, 129–131 (2007).
39. Lionel, N.D., Mirando, E.H., Nanayakkara, J.C. & Soysa, P.E. Levamisole in the treatment of ascariasis in children. *BMJ* **4**, 340–341 (1969).
40. Kaminsky, R. *et al.* Identification of the amino-acetonitrile derivative monepantel (AAD 1566) as a new anthelmintic drug development candidate. *Parasitol. Res.* **103**, 931–939 (2008).
41. Campbell, W.C., Fisher, M.H., Stapley, E.O., Albers-Schonberg, G. & Jacob, T.A. Ivermectin: a potent new antiparasitic agent. *Science* **221**, 823–828 (1983).
42. Hobert, O. The neuronal genome of *Caenorhabditis elegans*. *WormBook* **2013**, 1–106 (2013).
43. Richards, J.C., Behnke, J.M. & Duce, I.R. *In vitro* studies on the relative sensitivity to ivermectin of *Necator americanus* and *Ancylostoma ceylanicum*. *Int. J. Parasitol.* **25**, 1185–1191 (1995).
44. Geary, T.G. *et al.* Haemonchus contortus: ivermectin-induced paralysis of the pharynx. *Exp. Parasitol.* **77**, 88–96 (1993).
45. Bull, K. *et al.* Effects of the novel anthelmintic emodepside on the locomotion, egg-laying behaviour and development of *Caenorhabditis elegans*. *Int. J. Parasitol.* **37**, 627–636 (2007).
46. Willson, J., Amliwala, K., Harder, A., Holden-Dye, L. & Walker, R.J. The effect of the anthelmintic emodepside at the neuromuscular junction of the parasitic nematode *Ascaris suum*. *Parasitology* **126**, 79–86 (2003).
47. Zakon, H.H. Adaptive evolution of voltage-gated sodium channels: the first 800 million years. *Proc. Natl. Acad. Sci. USA* **109** (suppl. 1), 10619–10625 (2012).
48. Cohen, P. Protein kinases—the major drug targets of the twenty-first century? *Nat. Rev. Drug Discov.* **1**, 309–315 (2002).
49. Shah, N.P. *et al.* Overriding imatinib resistance with a novel ABL kinase inhibitor. *Science* **305**, 399–401 (2004).
50. Ashburn, T.T. & Thor, K.B. Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **3**, 673–683 (2004).
51. 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 Res.* **14**, 917–924 (2004).
52. Humphries, D.L. *et al.* The use of human faeces for fertilizer is associated with increased intensity of hookworm infection in Vietnamese women. *Trans. R. Soc. Trop. Med. Hyg.* **91**, 518–520 (1997).
53. Bethony, J. *et al.* Emerging patterns of hookworm infection: influence of aging on the intensity of Necator infection in Hainan Province, People's Republic of China. *Clin. Infect. Dis.* **35**, 1336–1344 (2002).
54. Quinnell, R.J. *et al.* Genetic and household determinants of predisposition to human hookworm infection in a Brazilian community. *J. Infect. Dis.* **202**, 954–961 (2010).
55. Miller, T.A. Vaccination against the canine hookworm diseases. *Adv. Parasitol.* **9**, 153–183 (1971).

© 2014 Nature America, Inc. All rights reserved.

ONLINE METHODS

Parasite material. The Anhui strain of *N. americanus* was maintained[56] in Golden Syrian Hamster (3–4 weeks, male) from Harlan under the George Washington University Institutional Animal Care and Use Committee–approved protocol 053-12.2, and in accordance with all Animal Welfare guidance. Adult worms were collected from intestines of hamsters infected subcutaneously with *N. americanus* iL3 for 8 weeks[57]. 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; environmental) and adult (parasitic) worm stages, were collected.

Sequencing, assembly and annotation. Fragment, paired-end whole-genome shotgun libraries (3 kb and 8 kb insert sizes) were sequenced using Roche/454 platform and assembled with Newbler[58]. A repeat library was generated (RepeatModeler) and repeats characterized (CENSOR[59] v. 4.2.27 against RepBase release 17.03 (ref. 60)). Ribosomal RNA genes (RNAmmer[61]) and transfer RNAs (tRNAscan-SE[62]) were identified. Other noncoding RNAs were identified by a sequence homology search against the Rfam database[63]. Repeats and predicted RNAs were then masked using RepeatMasker. Protein-coding genes were predicted using a combination of *ab initio* programs[64,65] and the annotation pipeline tool MAKER[66]. A consensus high-confidence gene set from the above prediction algorithms was generated (**Supplementary Note**). 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. These were then compared to the *C. elegans* genes (WS230). Significant differences in exon and intron lengths and numbers were tested between species and orthologous and nonorthologous gene groups using two-tailed *t*-tests with unequal variance (**Supplementary Note**). Two separate approaches were used to identify putative operons in *N. americanus* (**Supplementary Note**). Gene product naming was determined by BER (JCVI) and functional categories of deduced proteins were assigned[67–69]. Orthologous groups were built from 13 species using OrthoMCL[70], and genes not orthologous to the other 12 species were classified as *N. americanus* specific.

RNA sequencing. RNA was extracted[18], DNase treated and used to generate both Roche/454 and Illumina cDNA libraries (**Supplementary Note**) that were sequenced using a Genome Sequencer Titanium FLX (Roche Diagnostics) and Illumina (Illumina, San Diego, CA), with slight modifications (**Supplementary Note**). The 454 cDNA reads were analyzed as previously described[18]. The Illumina RNA-seq data were processed[71] and low-compositional complexity bases were masked[72]. RNA-seq reads were aligned[73] to the predicted gene set and genes with a breadth of coverage ≥50% across the gene sequence (i.e., "expressed") were used for further downstream analysis. Expression was quantified using expression values normalized to the depth of coverage per 100 million mapped bases (DCPM). Expressed genes were subject to further differential expression analysis using EdgeR[74] (false discovery rate <0.05) in order to identify stage-overexpressed genes (**Supplementary Note**).

Deduced proteome functional annotation and enrichment. Proteins were searched against KEGG[75] using KAAS[68] (cut-off 35 bits), and InterProScan[69] was used to get InterPro[76] domain matches and Gene Ontology[67] (GO) annotations. Proteins with signal peptides[77], nonclassical secretion[78] and transmembrane topology[77] were identified. The degradome was identified by comparison to the MEROPS[79] protease unit database using WU-BLAST (identifying the best hit with $E \leq e^{-10}$). 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[80]. GO enrichment significance comparing the iL3 and adult-overexpressed gene sets was calculated using FUNC[81] at a 0.01 significance threshold after Family-Wise Error Rate (FWER) population correction[81]. QuickGO[82] was used to analyze the hierarchical structure of significant GO categories.

Proteomic analysis of somatic worm extract. Whole worms were ground under liquid nitrogen before solubilization in lysis buffer, total protein was precipitated, and established methods[83] were used to reduce, alkylate and tryptic-digest two 1.5 mg samples of total somatic protein. Peptide fractions were prepared before LC and mass spectral analysis (**Supplementary Note**). Only proteins confirmed with at least two peptides and a confidence of $P \leq 0.05$ were considered identified. GO functional enrichment among the genes supported by proteomics was calculated[81], using all of the genes without proteomics support as a background for comparison.

Transcription factors and binding sites. Transcription factors in *N. americanus* were identified by extracting KEGG Orthology (KO) numbers from the KEGG transcription factor database (derived from TRANSFAC 7.0 (ref. 84)) and comparing to *N. americanus* KOs. Documented matrices of transcription factor binding sites were downloaded from the JASPAR database[85]. The corresponding protein accession numbers were extracted and converted to KOs, and were compared to *N. americanus* transcription factor KOs to define 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.

SCP/TAPS. Each protein was searched for the SCP/TAPS-representative protein domains[86] IPR014044 ("CAP domain") and PF00188 ("CAP")[86] using Interproscan[69] and hmmpfam[87]. Phylogenetic relationship trees using full length primary sequences derived from ungapped genes were built using Bayesian inference[88] and neighbor joining[89] as previously described for other helminths[32,86,90]. Leaves of the tree were annotated with domain information, secretion mode and expression data, and then visualized using iTOL[91].

Potential drug targets. GPCRs, LGICs and VGICs were identified with InterProScan[69]. Ion channels were identified using WU-BLASTP ($E \leq e^{-10}$) against the *C. elegans* proteome (WS230). For ivermectin target characterization, sequence alignments were obtained by MUSCLE[92] 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[93] using the *C. elegans* crystal structure as template[94]. 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[95]; protein structure models are rendered in PyMOL (Schrodinger, The PyMOL Molecular Graphics System, Version 1.3r1. 2010).

Kinome and chokepoints. *N. americanus* genes were screened against the collection of kinase domain models in the Kinomer[96], and custom score thresholds were applied for each kinase group and then adjusted until an hmmpfam search[87] came as close as possible to identifying known *C. elegans* kinases. Those same cutoffs were then applied to the *N. americanus* gene set to identify putative kinases as previously described[97]. Kinase prioritization was done by adapting the protocol previously described[10] (**Supplementary Note**).

Chokepoints of KEGG metabolic pathways were defined as a reaction that either consumes a unique substrate or produces a unique product. The reaction database from KEGG v58 (ref. 98) was used and the chokepoints were identified and prioritized as previously described[99] (**Supplementary Note**). Metabolic module abundances were calculated (and normalized in DCPM) based on KAAS annotations[68], and module bottlenecks were defined as reaction steps in the cascade that both are essential for the module completion and have low enzyme abundance that primarily constrains the overall module abundance. Homology models were aligned with their reference sequence using T-COFFEE[100], constructed with MODELLER[101] using default parameters and PDB structures with the highest sequence similarity, and docking was performed using AutoDock4.2 (ref. 102) using default parameters. Chemogenomic screening for compound prioritization was performed as previously described[99] (**Supplementary Note**).

Protein microarray. In 2005, 1494 individuals between the ages 4 and 66 years (inclusive) were enrolled (with informed consent) 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 Institutional Review Board (117040 and 060605), the Ethics Committee of Instituto René

© 2014 Nature America, Inc. All rights reserved.
npg

doi:10.1038/ng.2875

Rachou and the National Ethics Committee of Brazil (CONEP; protocol numbers 04/2008 and 12/2006). Venous blood (15 mL) was collected from individuals determined to be positive for *N. americanus* (**Supplementary Note**).

A total of 1,275 *N. americanus* open reading frames (ORFs) contained a classical signal peptide for secretion and had RNA-seq evidence for transcription in iL3 and/or adult worms. Of those, 623 corresponding cDNAs were successfully amplified, cloned, expressed and the protein extracts were contact-printed without purification onto nitrocellulose glass FAST slides (**Supplementary Note**). The printed *in vitro*–expressed proteins were quality-checked using antibodies against incorporated N-terminal polyhistidine (His) and C-terminal hemagglutinin (HA) tags.

Protein arrays were blocked in blocking solution (Whatman) and probed with human sera overnight. Arrays were washed, and isotype- and subclass-specific responses were detected using biotinylated mouse monoclonal antibodies against human IgG1 (Sigma, B6775, lot 031M4751, clone 8c/6-39), IgG3 (Sigma, B3523, lot 080M4811, clone HP-6050) and IgG4 (Sigma, B3648, lot 091M4783, clone HP-6025) and biotin-conjugated mouse monoclonal anti-human IgE Fc (Human Reagent Laboratory, Baltimore, MD, HP6061B). Microarrays were scanned using a GenePix microarray scanner (Molecular Devices). The data were analyzed using the "group average" method[103], whereby the mean fluorescence was considered for analysis (**Supplementary Note**).

© 2014 Nature America, Inc. All rights reserved.

56. 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**, 62–66 (2003).
57. Xiao, S. *et al*. The evaluation of recombinant hookworm antigens as vaccines in hamsters (*Mesocricetus auratus*) challenged with human hookworm, *Necator americanus*. *Exp. Parasitol.* **118**, 32–40 (2008).
58. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
59. 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).
60. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
61. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
62. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
63. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S.R. Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439–441 (2003).
64. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
65. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
66. Cantarel, B.L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
67. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
68. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185 (2007).
69. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).
70. Li, L., Stoeckert, C.J. Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
71. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
72. 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).
73. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
74. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
75. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
76. Hunter, S. *et al.* InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* **40**, D306–D312 (2012).
77. Käll, L., Krogh, A. & Sonnhammer, E.L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036 (2004).
78. Bendtsen, J.D., Jensen, L.J., Blom, N., Von Heijne, G. & Brunak, S. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.* **17**, 349–356 (2004).
79. Rawlings, N.D., Barrett, A.J. & Bateman, A. MEROPS: the peptidase database. *Nucleic Acids Res.* **38**, D227–D233 (2010).
80. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* **57**, 289–300 (1995).
81. Prüfer, K. *et al.* FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* **8**, 41 (2007).
82. Binns, D. *et al.* QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25**, 3045–3046 (2009).
83. Mulvenna, J. *et al.* Proteomics analysis of the excretory/secretory component of the blood-feeding stage of the hookworm, Ancylostoma caninum. *Mol. Cell Proteomics* **8**, 109–121 (2009).
84. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
85. 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 Res.* **36**, D102–D106 (2008).
86. 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).
87. Eddy, S.R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
88. Ronquist, F. & Huelsenbeck, J.P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
89. Larkin, M.A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
90. Cantacessi, C. *et al.* A portrait of the "SCP/TAPS" proteins of eukaryotes–developing a framework for fundamental research and biotechnological outcomes. *Biotechnol. Adv.* **27**, 376–388 (2009).
91. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).
92. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
93. Eswar, N. *et al.* Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* **50**, 2.9 (2007).
94. Hibbs, R.E. & Gouaux, E. Principles of activation and permeation in an anion-selective Cys-loop receptor. *Nature* **474**, 54–60 (2011).
95. 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**, 1189–1191 (2009).
96. Miranda-Saavedra, D. & Barton, G.J. Classification and functional annotation of eukaryotic protein kinases. *Proteins* **68**, 893–914 (2007).
97. Mitreva, M. *et al.* The draft genome of the parasitic nematode *Trichinella spiralis*. *Nat. Genet.* **43**, 228–235 (2011).
98. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **38**, D355–D360 (2010).
99. Taylor, C.M. *et al.* Discovery of anthelmintic drug targets and drugs using chokepoints in nematode metabolic pathways. *PLoS Pathog.* **9**, e1003505 (2013).
100. Notredame, C., Higgins, D.G. & Heringa, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
101. Sali, A. & Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
102. Morris, G.M. *et al.* AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).
103. Sundaresh, S. *et al.* Identification of humoral immune responses in protein microarrays using DNA microarray data analysis techniques. *Bioinformatics* **22**, 1760–1766 (2006).