nature genetics
OPEN

Genomics of *Loa loa*, a *Wolbachia*-free filarial parasite of humans

Christopher A Desjardins[1], Gustavo C Cerqueira[1], Jonathan M Goldberg[1], Julie C Dunning Hotopp[2], Brian J Haas[1], Jeremy Zucker[1], José M C Ribeiro[3], Sakina Saif[1], Joshua Z Levin[1], Lin Fan[1], Qiandong Zeng[1], Carsten Russ[1], Jennifer R Wortman[1], Doran L Fink[4,5], Bruce W Birren[1] & Thomas B Nutman[4]

***Loa loa*, the African eyeworm, is a major filarial pathogen of humans. Unlike most filariae, *L. loa* does not contain the obligate intracellular *Wolbachia* endosymbiont. We describe the 91.4-Mb genome of *L. loa* and that of the related filarial parasite *Wuchereria bancrofti* and predict 14,907 *L. loa* genes on the basis of microfilarial RNA sequencing. By comparing these genomes to that of another filarial parasite, *Brugia malayi*, and to those of several other nematodes, we demonstrate synteny among filariae but not with nonparasitic nematodes. The *L. loa* genome encodes many immunologically relevant genes, as well as protein kinases targeted by drugs currently approved for use in humans. Despite lacking *Wolbachia*, *L. loa* shows no new metabolic synthesis or transport capabilities compared to other filariae. These results suggest that the role of *Wolbachia* in filarial biology is more subtle than previously thought and reveal marked differences between parasitic and nonparasitic nematodes.**

Filarial nematodes dwell within the lymphatics and subcutaneous tissues of up to 170 million people worldwide and are responsible for notable morbidity, disability and socioeconomic loss[1]. Although eight filarial species infect humans, only five cause substantial pathology: *W. bancrofti*, *B. malayi* and *Brugia timori*, the causative agents of lymphatic filariasis; *Onchocerca volvulus*, the causative agent of 'river blindness' or onchocerciasis; and *L. loa*, the African eyeworm. *L. loa* affects an estimated 13 million people and causes chronic infection usually characterized by localized angioedema (Calabar swelling) and/or subconjunctival migration of adult worms across the eye ('African eyeworm'). Complications of infection include encephalopathy, entrapment neuropathy, glomerulonephritis and endomyocardial fibrosis[2]. *L. loa* is restricted geographically to equatorial west and central Africa, where its deerfly vector (*Chrysops* spp.) breeds. *L. loa* microfilariae (L1) are acquired by flies from human blood and subsequently develop into infective larvae (L3) before being reintroduced into a human host during a second blood meal (**Supplementary Fig. 1**). Although *L. loa* is the least studied of the pathogenic filariae, it has gained prominence recently because of the severe adverse events (encephalopathy and death) associated with ivermectin treatment[3] during mass drug administration campaigns in west and central Africa.

We targeted *L. loa* for genomic sequencing for two reasons. First, in contrast to other pathogenic filariae, *L. loa* lacks the α-proteobacterial endosymbiont *Wolbachia*. The obligate nature of *Wolbachia* symbiosis in *W. bancrofti*, *B. malayi* and *O. volvulus* has been inferred by studies in which antibiotics (for example, doxycycline) that target *Wolbachia* (but not the worm itself) have shown efficacy in treating humans with these infections[4,5]. Through genomic analysis, *Wolbachia* have been hypothesized to provide essential metabolic supplementation to their filarial hosts[6,7]. The absence of the *Wolbachia* endosymbiont in *L. loa* suggests that either there has been lateral transfer of important bacterially encoded genes or the obligate relationship between the endosymbiont and its filarial host is dispensable, at least under certain circumstances. Understanding the comparable adaptations of *L. loa* is considered essential to gain insight into the potential impact of the endosymbiont[8]. Second, as the most neglected of the pathogenic filariae, but one gaining clinical prominence, understanding the host-parasite relationship as it relates to the severe post-treatment reactions typical of both *Wolbachia*-containing and *Wolbachia*-free filarial parasites is of paramount importance.

Thus, we generated a draft genome sequence of *L. loa* and produced a refined gene annotation aided by transcriptional data from *L. loa* microfilariae. We also generated draft genome sequences of two of the most pathogenic (and *Wolbachia*-containing) filarial species, *W. bancrofti* and *O. volvulus*. This approach enabled us to more comprehensively define the genomic differences between *L. loa* and other filarial parasites.

RESULTS

Genome assemblies and repeat content

The nuclear genome of *L. loa* consists of five autosomes plus a sex chromosome. Using 454 whole-genome shotgun sequencing, we sequenced *L. loa* to 20× coverage and assembled it into 5,774 scaffolds

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Institute for Genome Science, Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, Maryland, USA. [3]Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland, USA. [4]Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland, USA. [5]Present address: Center for Biologics Evaluation and Research, Food and Drug Administration, Rockville, Maryland, USA. Correspondence should be addressed to T.B.N. (tnutman@niaid.nih.gov).

Received 2 November 2012; accepted 22 February 2013; published online 24 March 2013; doi:10.1038/ng.2585

Table 1 Genome features of filarial worms and their *Wolbachia* endosymbionts

Organism	Coverage	Sequence (Mb)	Scaffolds	Scaffold N50 (kb)	GC (%)	Repetitive (%)	Low complexity (%)	Genes (*n*)
L. loa	20×	91.4	5,774	172	31.0	9.3	1.7	14,907[a]
W. bancrofti	12×	81.5	25,884	5.16	29.7	6.2	3.9	19,327[a]
O. volvulus	5×	26.0	22,675	1.27	32.5	–	–	–
B. malayi	9×	93.7	8,180	94	30.2	12.1	1.1	18,348
wBm	11×	1.08	1	–	34.2	–	–	805
wWb	2×	1.05	763	1.62	34.0	–	–	–
wOv	2×	0.44	341	1.51	32.8	–	–	–

[a]Because of fragmentation of the genome assemblies, the true *W. bancrofti* gene count is estimated to be 14,496–15,075 genes, whereas the true *L. loa* gene count is estimated to be 14,261 genes (**Supplementary Note**). *wBm*, *Wolbachia* of *B. malayi*; *wWb*, *Wolbachia* of *W. bancrofti*; *wOv*, *Wolbachia* of *O. volvulus*.

with an N50 of 172 kb and total size of 91.4 Mb (**Table 1**). We sequenced the *W. bancrofti* and *O. volvulus* genomes derived from single adult worms (an unsexed juvenile adult worm for *W. bancrofti* and an adult male worm for *O. volvulus*) to 12× and 5× coverage, respectively (**Table 1**). Because of the low coverage of the *O. volvulus* genome, we did not include it in further analyses. Although the assembly sizes of the *L. loa* and *B. malayi* genomes are comparable (91.4 Mb and 93.7 Mb, respectively), the scaffold N50 of the *L. loa* genome is almost twice that of the *B. malayi* genome, making the *L. loa* genome assembly the most contiguous of any filarial nematode so far. The filarial genomes differ widely in repeat content (**Table 1**, **Supplementary Tables 1–14** and **Supplementary Note**), with the *L. loa* genome being more repetitive than that of *W. bancrofti* but less repetitive than that of *B. malayi*.

As nuclear *Wolbachia* transfers (nuwts) have been identified in all *Wolbachia*-colonized and *Wolbachia*-free filarial nematodes examined so far[9], we expected to find similar transfers in the *L. loa* genome. However, a BLAST-based search of the assembled *L. loa* genome did not reveal any large transfers of *Wolbachia* DNA. A more sensitive read-based analysis determined that the *L. loa* genome does not have any large (>500 bp), recent transfers (**Supplementary Note**). It does however have small, presumably older transfers, supporting the hypothesis that *L. loa* was once colonized by *Wolbachia* but subsequently lost its endosymbiont (**Supplementary Table 15** and **Supplementary Fig. 2**). Of the transfers that are definitively of *Wolbachia* ancestry and not of possible mitochondrial ancestry, there is no evidence that they are functional in *L. loa* (**Supplementary Note**).

Gene content and synteny

We produced initial gene sets for both *L. loa* and *W. bancrofti* using a combination of gene predictors with refinements to the *L. loa* annotation on the basis of RNA sequencing (RNA-Seq) data (Online Methods). The final *L. loa* gene set contained 14,907 genes, 70% of which were supported by RNA-Seq (**Table 1** and **Supplementary Tables 16** and **17**). The *W. bancrofti* genome is predicted to encode 19,327 genes (**Table 1** and **Supplementary Note**). The filarial genomes showed a high degree of synteny (**Fig. 1**), with 40% and 13% of *L. loa* genes being syntenic relative to *B. malayi* and *W. bancrofti*, respectively. Nearly all the syntenic breaks between filarial genomes occurred at scaffold ends (**Supplementary Fig. 3b**), suggesting that the synteny percentage detected was limited by assembly contiguity and the true level of synteny is much higher. When we compared the *L. loa* genome to that of *Caenorhabditis elegans*, orthologs from

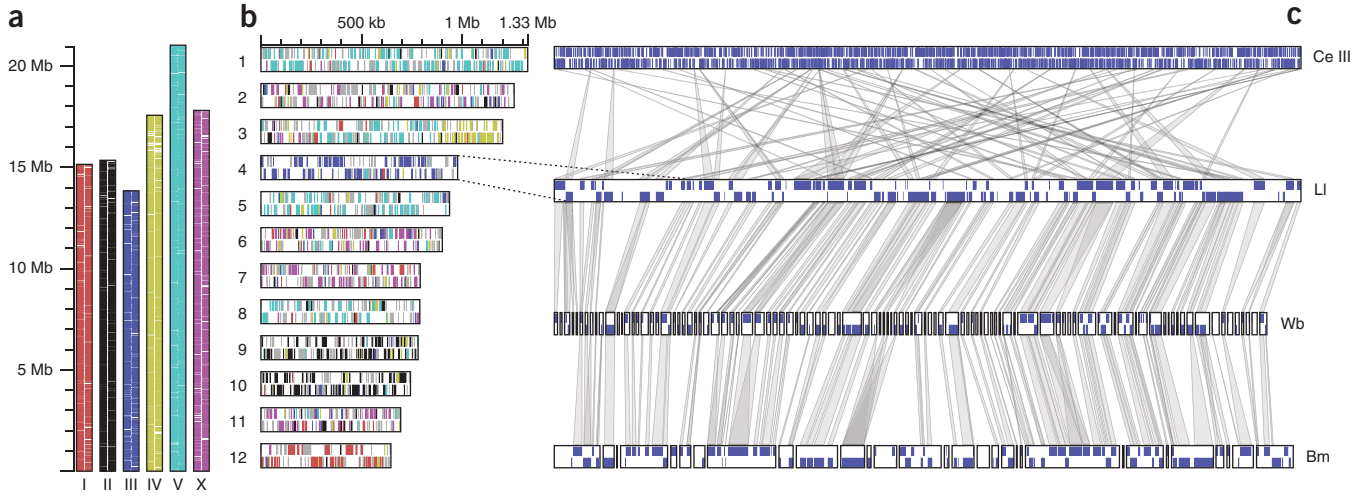

Figure 1 Synteny between filarial worms and *C. elegans*. (**a**) Gene distribution on the *C. elegans* genome. Vertical bars represent the six *C. elegans* chromosomes (labeled at the bottom). Horizontal boxes within each bar indicate the location and strand of *C. elegans* genes (boxes on the left indicate the plus strand, and boxes on the right indicate the minus strand). (**b**) Gene distribution on the 12 longest *L. loa* scaffolds. Scaffolds are represented by horizontal bars and identified by labels on the left. Vertical colored boxes indicate the position and strand of each gene (boxes on the top indicate the plus strand, and boxes on the bottom indicate the minus strand). The color coding indicates the chromosome on which each ortholog in *C. elegans* is located and is consistent with the colors used in **a**. Gray boxes represent either genes without orthologs in *C. elegans* or genes with two or more homologs in distinct *C. elegans* chromosomes. (**c**) Distribution of *L. loa* (LI) scaffold 4 orthologs on the *C. elegans* chromosome 3 (Ce III), *W. bancrofti* (Wb) and *B. malayi* (Bm) genomes. The scaffolds and chromosomes with the best matches to *L. loa* scaffold 4 on the basis of whole-genome alignment are shown. Each row contains one or more horizontal bars representing either chromosomes (*C. elegans*) or scaffolds (*L. loa*, *B. malayi* and *W. bancrofti*) from each sequenced genome. Purple boxes indicate the position and strand of the genes. Gray projections connect orthologous genes across organisms.

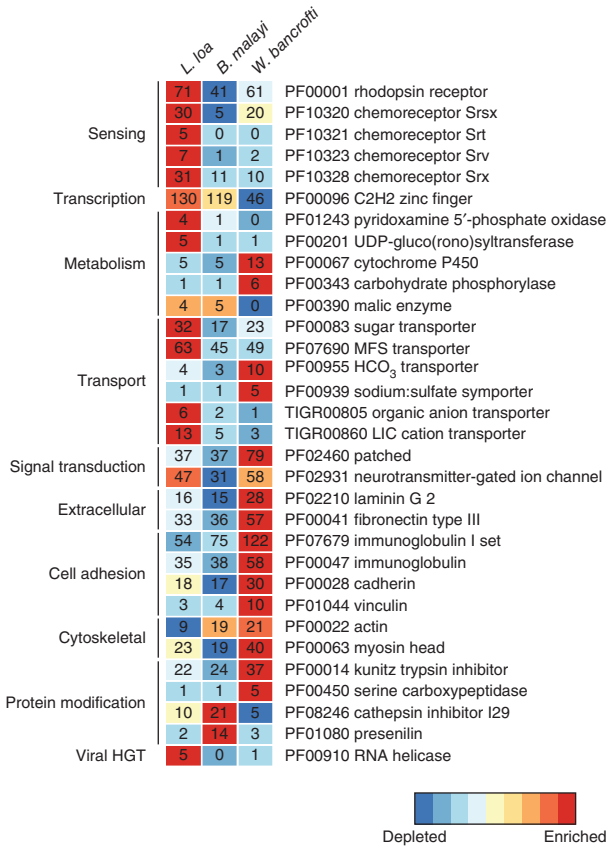

Figure 2 Enriched and depleted Pfam and TIGRFAM domains in each filarial genome relative to the other two. All domains significantly enriched ($P < 0.05$, Fisher's exact test) are shown; red indicates enriched, and blue indicates depleted. The numbers of domains identified are shown in each box. Broad functional categories representing each domain are shown to the left.

a single *L. loa* scaffold mapped predominantly to a single *C. elegans* chromosome (**Fig. 1**). However, only 2% of all *L. loa* genes were syntenic relative to *C. elegans* (**Supplementary Fig. 3a**), supporting the hypothesis that genome rearrangements during filarial evolution were mostly intrachromosomal[7]. There was an intermediate level of synteny (12%) between *L. loa* and the related nonfilarial parasite *Ascaris suum* (**Supplementary Fig. 3a**).

We were able to assign more than half of the genes encoded by the *L. loa* and *W. bancrofti* genomes to functional categories, Pfam domains, Gene Ontology (GO) terms and/or Enzyme Commission (EC) numbers (**Supplementary Fig. 4** and **Supplementary Tables 16** and **18**). Relative to other filarial genomes, the *L. loa* genome is enriched ($P < 0.05$, Fisher's exact test) for numerous domains, including that containing pyridoxamine 5′-phosphate oxidases that synthesize vitamin B (**Fig. 2**). The *L. loa* genome is also enriched for numerous chemoreceptors, suggesting that *L. loa* may be capable of more complex interactions with its host environment than are other filarial worms (**Supplementary Note**). An RNA helicase domain involved in viral DNA replication is enriched in the *L. loa* genome; this domain was probably horizontally transferred to the *L. loa* genome from cyclovirus infection (**Supplementary Note**). Although not statistically significant, the *L. loa* genome encodes more hyaluronidases (six) than the *B. malayi* or *W. bancrofti* genomes (two each). Hyaluronidases are often involved in tissue penetration and could allow *L. loa* to move more readily through human host tissue, as *L. loa* adults are highly mobile, whereas *B. malayi* and *W. bancrofti* adults are commonly tethered to the lymphatic endothelium.

The genome of *W. bancrofti* is enriched ($P < 0.05$, Fisher's exact test) for genes with domains related to cellular adhesion and the extracellular matrix (for example, cadherins, laminins and fibronectins). Whether these are important in mediating the fibrosis associated with lymphatic filarial disease (for example, elephantiasis or lymphedema) in *W. bancrofti* infection[10] or with establishing an anatomical niche within the afferent lymphatics where the adults reside awaits clarification.

Gene products associated with immunologic responses

Each filarial parasite interacts with both its definitive mammalian host and its intermediate arthropod host (*Chrysops* spp. in the case of *L. loa*) during its life cycle (**Supplementary Fig. 1**). The parasite is thought not only to have its own innate immune system to protect itself from microbial pathogens but also to have evolved mechanisms to exploit and/or subvert host and vector defense mechanisms. Although adaptive immune molecules such as immunoglobulins or Toll-like receptors (TLRs) are absent in *L. loa* and other filarial nematodes, *L. loa*, similarly to other filariae, seems to have a primordial Toll-related pathway (**Supplementary Table 19** and **Supplementary Note**). The innate immune system encoded by the *L. loa* genome also includes C-type lectins, galectins, jacalins and scavenger receptors. *L. loa* contains a number of lipopolysaccharide binding proteins that have been implicated in modulating the effects of host bacteria or microbial translocation products. Similarly to *B. malayi*, the *L. loa* and *W. bancrofti* genomes do not encode antibacterial peptides described in *C. elegans* and *A. suum*[7], suggesting that these molecules are either dispensable in filariae or too divergent to detect.

Analysis of *L. loa* genes identified a number of human cytokine and chemokine mimics and/or antagonists, including genes encoding macrophage migration inhibition factor (MIF) family signaling molecules, transforming growth factor-β and their receptors, members of the interleukin-16 (IL-16) family, an IL-5 receptor antagonist, an interferon regulatory factor, a homolog of suppressor of cytokine signaling 7 (SOCS7) and two members of the chemokine-like family (**Supplementary Table 19**). In addition, the *L. loa* genome encodes 17 serpins and 7 cystatins, which have been shown to interfere with antigen processing and presentation to T cells[11], 2 indoleamine 2,3-dioxygenase (IDO) genes, which encode immunomodulatory proteins implicated in strategies of immune subversion, and a number of members of the Wnt family of developmental regulators, which typically modulate immune activation. The *L. loa* genome encodes proteins that have sequences similar to those of human autoantigens (**Supplementary Note**). Although some of these putative autoantigens can also be found in the other filariae, the slight expansion of them in *L. loa* suggests that antibodies induced by *L. loa* infection may be more autoreactive than those induced by other parasites.

Protein kinases

In addition to elucidating host-pathogen interactions, pathogen genomes can be evaluated for potential drug targets, such as protein kinases. We therefore annotated protein kinases in the *L. loa* genome and compared them to those in other nematode genomes (**Supplementary Tables 20–23** and **Supplementary Fig. 5**). We found numerous differences between filarial and nonparasitic nematode kinases, particularly regarding those involved in meiosis. The widely conserved TTK kinase (MPS1), which has a key role in eukaryotic meiosis[12], is present in *L. loa* and absent in *C. elegans*. By contrast, filarial nematodes lack the nearly universally conserved RAD53-family

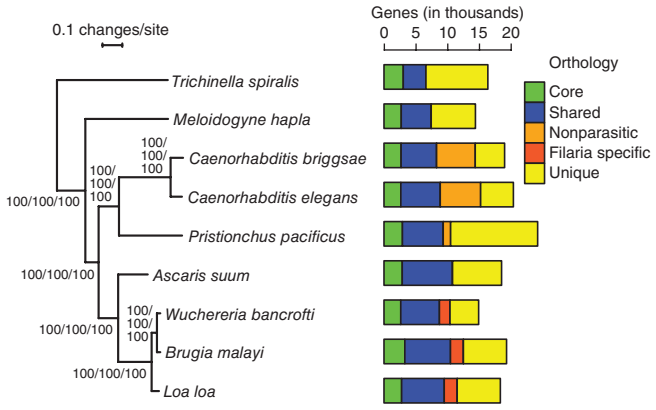

Figure 3 Phylogenomic analysis of nematodes. Maximum likelihood, parsimony and Bayesian methods all estimated an identical phylogeny using the concatenated protein sequences of 921 single-copy orthologs. To the left of each node are likelihood bootstrap support values/parsimony bootstrap support values/Bayesian posterior probabilities. The distributions of genes in the ortholog clusters are shown to the right of the phylogeny. Core genes are encoded by all genomes, shared genes are encoded by at least two but fewer than all genomes, and unique genes are found only in one genome. Orthologs specific to the nonparasitic (*C. elegans*, *C. briggsae* and *P. pacificus*) and filarial nematodes are also highlighted. Of the 6,280 *L. loa* genes with no functional assignment, 3,665 are unique to *L. loa* and 1,158 are filarial specific.

kinase CHK-2, which is present in *C. elegans*. In most eukaryotes, RAD53 is involved in initiating cell-cycle arrest when DNA damage is detected, but in *C. elegans* it is essential for chromosome synapsis and nuclear rearrangement during meiosis[13]. This reciprocal difference suggests that meiosis in filarial parasites may be regulated in a manner more similar to that in typical eukaryotes than in *C. elegans* (**Supplementary Note**). Six *L. loa* protein kinases are orthologous to targets of drugs currently approved for use in humans (**Supplementary Table 23**), including the tyrosine kinase inhibitor imatinib, which has been shown to kill schistosomes[14] and *Brugia* parasites of all stages at concentrations ranging from 5 to 50 μM (T.B.N., unpublished data). Therefore, repurposing already approved drugs that target these kinases may be promising in treating filarial (and other helminth) infections[15].

Nematode phylogenomics

To examine the evolution of filarial parasites in the context of other nematodes, we estimated a phylogeny from 921 single-copy core orthologs across nine nematode genomes using maximum likelihood, parsimony and Bayesian methods. All methods converged on a single topology with 100% support (either bootstrap values or posterior probabilities) at all nodes (**Fig. 3**). This phylogeny indicates that *Meloidogyne hapla* occupies a position basal to a clade of Rhabditina (*C. elegans*, *Caenorhabditis briggsae* and *Pristionchus pacificus*) and the Spirurina (filarial worms and *A. suum*). Although these results contrast with previous studies based on ribosomal subunits that placed *M. hapla* closer to Rhabditina than to the filarial worms[16,17], our analysis used a larger gene set and had higher nodal support values.

Relative to the genomes of nonparasitic nematodes, we identified numerous orthologs as being unique to the filarial parasites (**Fig. 3**). Proteins encoded by the filarial genomes showed enrichment of immunogenic domains such as extracellular and cell-adhesion domains and in a metabolic context were enriched for trehalase domains involved in trehalose degradation ($q < 0.05$, Fisher's exact test; **Supplementary Fig. 6**). Trehalose is known to be involved in the protection of nematodes from environmental stress[18] and could potentially have a key role in filarial survival. Trehalose and its biosynthetic pathway have been shown to be associated with increased lifespan in *C. elegans*[19] and might support the idea that increased use of trehalose by filarial nematodes could be related to their relatively long lifespan.

The filarial genomes lack a wide array of seven-transmembrane G protein–coupled chemoreceptors (7TM GPCRs; **Supplementary Fig. 6**). Profiling of 7TM GPCRs revealed a pattern of progressive loss of many families in the transition from nonparasitic to parasitic lifestyles (**Fig. 4**). For example, filarial nematodes and *Trichinella spiralis* completely lack the STR superfamily, including ODR-10, which is known to be involved in detection of volatiles[20], and KIN-29, a protein kinase that regulates STR expression in *C. elegans*[21]. If the STR superfamily is more broadly involved in odorant detection, this could explain why these molecules are lacking in filarial nematodes and *T. spiralis* parasites that live only in aqueous environments, whereas they are retained in *A. suum* and *M. hapla*, which are exposed to volatiles in part of their life cycle. Only the SRAB, SRX, SRSX and SRW families were conserved across all nematodes, suggesting that these 7TM GPCRs mediate vital nematode functions.

Filarial genomes are also depleted in both soluble and receptor guanylate cyclases; these cyclases are involved in the regulation of environmental sensing and complex sensory integration functions (**Fig. 4**). However, GCY-35 and GCY-36, which are involved in the detection of molecular oxygen in solution[22], are encoded in the filarial genomes. Protein kinase profiling revealed 18 receptor guanylate cyclases that are present in *C. elegans* but not in filarial worms, including the environmental sensors GCY-14 and GCY-22 (**Supplementary Table 23**). Depletion of these and other kinases involved in olfactory and gustatory sensing, including KIN-29, suggests that the environments of filarial nematodes are less complex in terms of chemosensory inputs than are those inhabited by nonparasitic nematodes (**Supplementary Note**). The *L. loa* genome does, however, encode significantly more chemoreceptors than do other

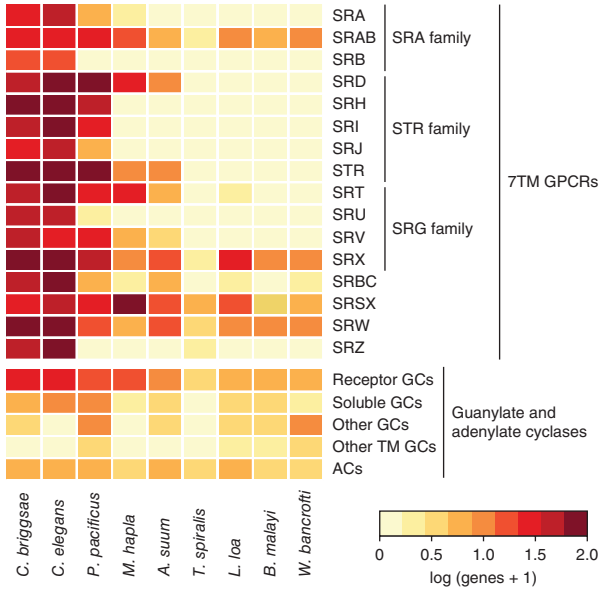

Figure 4 Phylogenetic profile of chemoreceptors in nematode genomes. The 7TM GPCRs, guanylate cyclases (GCs) and adenylate cyclases (ACs) are shown. TM, transmembrane.

Table 2 Phylogenetic profiles of biosynthesis pathways hypothesized to be involved in filaria-*Wolbachia* symbiosis

Biosynthesis pathway	*C. elegans*	*C. briggsae*	*P. pacificus*	*M. hapla*	*T. spiralis*	*A. suum*	*B. malayi*	*W. bancrofti*	*L. loa*	*wBm*	*wMel*	*wPip*	*wWb*
Heme	–	–	–	–	–	–	–[a]	–[a]	–[a]	+	+	+	+
Riboflavin	–	–	–	–	–	–	–	–	–	+	+	+	+
FAD	+	+	+/–	+/–	+	+	+	+	+	+	+	+	+/–
Glutathione	+	+	+	+	+	+	+	+	+	+	+	+	+
Purines	+	+	–	–	+/–	+	–	–	–	+	+	+	+
Pyrimidines	+	+	+	+	+/–	+	+/–	+/–	+/–	+	+	+	+

[a]All filarial worms encode a ferrochelatase, the last enzyme in heme synthesis (**Supplementary Note**). The conservation of each pathway across nematodes and across *Wolbachia* are shown in **Supplementary Tables 24** and **25**, respectively. Pathways are labeled as complete (+), partial (+/–) or absent (–). *wBm*, *Wolbachia* of *B. malayi*; *wMel*, *Wolbachia* of *D. melanogaster*; *wPip*, *Wolbachia* of *Culex pipiens*; *wWb*, *Wolbachia* of *W. bancrofti*.

filarial nematodes ($P < 0.05$, Fisher's exact test), which may be related to the increased mobility of *L. loa* adult worms.

Phylogenetic profiling of metabolism

Previous genomic analysis identified five biosynthetic pathways (heme, riboflavin, FAD, glutathione and nucleotide synthesis) present in *Wolbachia* but missing from its relatives, for example, *Rickettsia*. These *Wolbachia*-encoded pathways were hypothesized to provide metabolites needed by their filarial hosts[6]. As *L. loa* lacks *Wolbachia*, it was theorized that the *L. loa* genome must encode genes to replace these pathways, potentially laterally transferred from *Wolbachia* to an ancestor of *L. loa*. However, no transfers relating to these metabolic functions were apparent. Thus, we generated complete metabolic pathway reconstructions for nine nematode and four *Wolbachia* genomes (**Table 2** and **Supplementary Tables 24** and **25**) to determine how *L. loa* acquires these metabolites and placed the results in an evolutionary context. None of the five 'complementary' pathways differed between *L. loa* and the other filarial nematodes, calling into question the role of these pathways in filarial-*Wolbachia* symbiosis.

Furthermore, in only two pathways (heme and nucleotide synthesis) did the filarial genomes differ from those of the other nematodes. The FAD and glutathione pathways are complete in all nematode genomes, whereas the riboflavin pathway is missing from all nematode genomes. The heme biosynthesis pathway, previously reported to be absent in *B. malayi*[7], is missing from not only the filarial worms but also all nematode genomes characterized so far. Experimental work on *C. elegans* (which is also *Wolbachia* free) has shown that it cannot synthesize heme *de novo*[23]. *B. malayi* has been previously noted as having a single member of the heme synthesis pathway, ferrochelatase (an enzyme that catalyzes the last step in heme synthesis[7]; **Supplementary Note**). The gene encoding ferrochelatase is also present in the *L. loa* and *W. bancrofti* genomes but is absent in all other nematode genomes, including that of *A. suum*. It is possible that this gene in filarial nematodes is not involved in heme synthesis but rather in an alternate, unknown pathway.

Similarly to *B. malayi*, both *L. loa* and *W. bancrofti* lack the ability to synthesize nucleotides *de novo*. All three filarial genomes lack the majority of the proteins involved in the purine synthesis pathway, as well as the first enzyme involved in the pyrimidine synthesis pathway (**Table 2** and **Supplementary Table 24**). Other nematodes have also lost portions of these pathways; the purine synthesis pathway has been largely lost in *P. pacificus* and *M. hapla*, whereas the first two enzymes in the pyrimidine synthesis pathway have been lost in *T. spiralis*. These multiple and probably independent losses could underscore a general flexibility in the need for *de novo* nucleotide synthesis in nematodes. All nematodes, including the filariae, have complete sets of purine and pyrimidine interconversion pathways (**Supplementary Table 24**), implying that they could generate all necessary nucleotides from a single purine and pyrimidine source, a concept supported by experimental data in *B. malayi*[24]. Filarial genomes encode two purine-specific 5′ nucleotidases for salvage, whereas all other nematodes encode only one; the extra copy in the filariae seems to have arisen from a single gene duplication event and diverged markedly from the ancestral gene (**Supplementary Fig. 7**). Additionally, we profiled known nematode and *Wolbachia* transporters linked to these pathways and found no evidence of differences between filarial and nonfilarial nematodes or among *Wolbachia* endosymbionts (**Supplementary Note** and **Supplementary Fig. 8**). Given the uniformity of these pathways across nematodes and the apparent lack of any related transfers of *Wolbachia* DNA to the *L. loa* genome, it is probable that the symbiotic role of *Wolbachia* in filarial nematodes either lies outside these pathways or involves more subtle metabolic supplementation rather than the wholesale provision of unproduced metabolites.

The only metabolic pathway found to differ in gene content between *L. loa* and other nematodes with sequenced genomes is vitamin B6 synthesis and salvage. Most nematode genomes encode single copies of the two enzymes involved in vitamin B6 salvage, but the *L. loa* genome encodes five copies of the second enzyme, pyridoxal 5′-phosphate synthase (**Supplementary Note**). This pathway also differed among *Wolbachia* genomes. Although both of the insect *Wolbachia* genomes also encoded two genes involved in the synthesis of vitamin B6 (*pdxJ* and *pdxK*), neither of the filarial *Wolbachia* genomes did (the difference between *Wolbachia* of *B. malayi* and *Wolbachia* of *Drosophila melanogaster* was noted previously[6]). If the filarial *Wolbachia* endosymbionts need to acquire vitamin B6 exogenously, this could explain a metabolic need of *Wolbachia* that is fulfilled by the nematode. However, with that hypothesis in mind, it is unclear why *L. loa*, the one pathogenic filarial nematode without *Wolbachia*, would encode a greater number of vitamin B6 salvage genes than either *B. malayi* or *W. bancrofti*. We could not exclude differences in pyridoxine transporters, as we could identify no orthologs of known transporters in either nematode or *Wolbachia* genomes (**Supplementary Note**).

DISCUSSION

The study of some nematode genomes has already provided great insight into the genomic structure, biology and evolution of this major division of nematode parasites. With the release of the genome of *L. loa*, a human pathogen and parasitic nematode that does not contain *Wolbachia*, we have been able provide insights into the dispensability of this endosymbiont that deepen the mystery surrounding the 'essential nature' of *Wolbachia* for many filarial worms.

Through large-scale genomic comparisons within the phylum Nematoda, we have not only been able to define molecules and pathways that are either *L. loa*–specific or filaria-specific but also, by comparison with nonparasitic nematodes (for example, *C. elegans*), gained a glimpse into the nature of parasitism itself. Moreover, this

effort has identified new targets for intervention that should aid programs aimed at the control and elimination of these important but neglected parasites.

URLs. Repeat masker, http://www.repeatmasker.org; GLU package, http://code.google.com/p/glu-genetics; TransposonPSI, http://transposonpsi.sourceforge.net; Pristionchus database, http://www.pristionchus.org; WormBase, http://www.wormbase.org; KinBase database, http://www.kinase.com; WormCyc database, http://wormcyc.broadinstitute.org.

METHODS

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

Accession codes. All genome assemblies are available in GenBank under the following BioProject identifiers and accession numbers, respectively: *L. loa* (PRJNA37757 and ADBU02000000), *W. bancrofti* (PRJNA37759 and ADBV01000000), *O. volvulus* (PRJNA37761 and ADBW01000000), *Wolbachia* of *W. bancrofti* (PRJNA43539 and ADHD00000000) and *Wolbachia* of *O. volvulus* (PRJNA43537 and ADHE00000000).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank members of the Broad Institute Genomics Platform for sequencing and D. Neafsey for comments on the manuscript. This project has been funded in part by the National Institute of Allergy and Infectious Diseases, US National Institutes of Health (NIH), Department of Health and Human Services under contract number HHSN272200900018C and by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, NIH. J.C.D.H. is funded by the NIH Director's New Innovator Award Program (1-DP2-OD007372).

AUTHOR CONTRIBUTIONS

T.B.N., B.W.B. and D.L.F. conceived and designed the project. T.B.N. and D.L.F. provided the samples. J.Z.L., L.F. and C.R. coordinated and/or conducted the sequencing. S.S. assembled the genomes. J.M.G., B.J.H. and Q.Z. annotated the genomes. C.A.D., T.B.N., G.C.C., J.M.G., J.C.D.H., J.Z., D.L.F. and J.M.C.R. analyzed the genomes. C.A.D., T.B.N., G.C.C., J.M.G. and J.C.D.H. wrote the paper. T.B.N., B.W.B., J.R.W. and B.J.H. supervised and coordinated the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Hotez, P.J. *et al.* Control of neglected tropical diseases. *N. Engl. J. Med.* **357**, 1018–1027 (2007).
2. Klion, A.D. & Nutman, T.B. in *Tropical Infectious Diseases: Principles, Pathogens and Practice* (eds. Guerrant, R.L., Walker, D.H. & Weller, P.F.) 735–740 (Churchill Livingstone, 2011).
3. Gardon, J. *et al.* Serious reactions after mass treatment of onchocerciasis with ivermectin in an area endemic for *Loa loa* infection. *Lancet* **350**, 18–22 (1997).
4. Taylor, M.J., Hoerauf, A. & Bockarie, M. Lymphatic filariasis and onchocerciasis. *Lancet* **376**, 1175–1185 (2010).
5. Coulibaly, Y.I. *et al.* A randomized trial of doxycycline for *Mansonella perstans* infection. *N. Engl. J. Med.* **361**, 1448–1458 (2009).
6. Foster, J. *et al.* The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS Biol.* **3**, e121 (2005).
7. Ghedin, E. *et al.* Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317**, 1756–1760 (2007).
8. Saint André, A. *et al.* The role of endosymbiotic *Wolbachia* bacteria in the pathogenesis of river blindness. *Science* **295**, 1892–1895 (2002).
9. Dunning Hotopp, J.C. Horizontal gene transfer between bacteria and animals. *Trends Genet.* **27**, 157–163 (2011).
10. Anuradha, R. *et al.* Altered circulating levels of matrix metalloproteinases and inhibitors associated with elevated type 2 cytokines in lymphatic filarial disease. *PLoS Negl. Trop. Dis.* **6**, e1681 (2012).
11. Hartmann, S. & Lucius, R. Modulation of host immune responses by nematode cystatins. *Int. J. Parasitol.* **33**, 1291–1302 (2003).
12. Gilliland, W.D. *et al.* The multiple roles of mps1 in *Drosophila* female meiosis. *PLoS Genet.* **3**, e113 (2007).
13. Meier, B. & Ahmed, S. Checkpoints: chromosome pairing takes an unexpected twist. *Curr. Biol.* **11**, R865–R868 (2001).
14. Beckmann, S. & Grevelding, C.G. Imatinib has a fatal impact on morphology, pairing stability and survival of adult *Schistosoma mansoni in vitro*. *Int. J. Parasitol.* **40**, 521–526 (2010).
15. Dissous, C. & Grevelding, C.G. Piggy-backing the concept of cancer drugs for schistosomiasis treatment: a tangible perspective? *Trends Parasitol.* **27**, 59–66 (2011).
16. Meldal, B.H. *et al.* An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa. *Mol. Phylogenet. Evol.* **42**, 622–636 (2007).
17. Blaxter, M.L. *et al.* A molecular evolutionary framework for the phylum Nematoda. *Nature* **392**, 71–75 (1998).
18. Pellerone, F.I. *et al.* Trehalose metabolism genes in *Caenorhabditis elegans* and filarial nematodes. *Int. J. Parasitol.* **33**, 1195–1206 (2003).
19. Honda, Y., Tanaka, M. & Honda, S. Trehalose extends longevity in the nematode *Caenorhabditis elegans*. *Aging Cell* **9**, 558–569 (2010).
20. Sengupta, P., Chou, J.H. & Bargmann, C.I. *odr-10* encodes a seven transmembrane domain olfactory receptor required for responses to the odorant diacetyl. *Cell* **84**, 899–909 (1996).
21. van der Linden, A.M. *et al.* The EGL-4 PKG acts with KIN-29 salt-inducible kinase and protein kinase A to regulate chemoreceptor gene expression and sensory behaviors in *Caenorhabditis elegans*. *Genetics* **180**, 1475–1491 (2008).
22. Zimmer, M. *et al.* Neurons detect increases and decreases in oxygen levels using distinct guanylate cyclases. *Neuron* **61**, 865–879 (2009).
23. Rao, A.U., Carta, L.K., Lesuisse, E. & Hamza, I. Lack of heme synthesis in a free-living eukaryote. *Proc. Natl. Acad. Sci. USA* **102**, 4270–4275 (2005).
24. Rajan, T.V. Exogenous nucleosides are required for the morphogenesis of the human filarial parasite *Brugia malayi*. *J. Parasitol.* **90**, 1184–1185 (2004).

ONLINE METHODS

Sequencing and assembly. For *L. loa*, 5×10^5 microfilariae were purified during a therapeutic apheresis from a patient with loiasis infected in Cameroon seen at the NIH under protocol 88-I-83 (NCT00001230). A single unfertilized adult *W. bancrofti* worm was obtained under ultrasonic guidance (as part of protocol NCT00339417) in Tieneguebougou, Mali. A single adult *O. volvulus* male was isolated from a surgically removed subcutaneous nodule in Ecuador after collagenase digestion. Genomic DNA for all samples was prepared using the Qiagen genomic DNA kit (Qiagen, Gaithersburg, MD). DNA obtained from *W. bancrofti* and *O. volvulus* was amplified using the Qiagen Repli-g Midi Kit. For *L. loa*, *W. bancrofti* and *O. volvulus*, approximately 50, 10 and 5 µg of DNA, respectively, was used for genomic sequencing. For *L. loa*, 454 shotgun fragment and 3-kb jumping sequencing libraries were prepared and sequenced as previously described[25]. Only fragment libraries were constructed for *W. bancrofti* and *O. volvulus*. Assemblies were then generated using Newbler version 2.1 (Roche 454 Life Sciences). Given the overall low coverage of the *W. bancrofti* and *O. volvulus* assemblies (5×–12×), no bias normalization was done for the whole-genome amplified sequence data. Also for the *W. bancrofti* and *O. volvulus* assemblies, contigs were screened by BLASTing against GenBank's nonredundant nucleotide database (NT) using a cutoff of 1×10^{-25} and minimum match length of 100 bp, and all contigs where the top match was to *Wolbachia* were removed. Any contigs remaining in the nematode assembly that had secondary matches to *Wolbachia* were screened manually to ensure that no large chimeric contigs had been generated and retained. Unassembled reads were also screened for the *Wolbachia* sequence using the same BLAST parameters and database. Unassembled reads identified as *Wolbachia*, along with reads underlying the contigs identified as *Wolbachia*, were assembled together using Newbler version 2.1 to generate the *Wolbachia* genome assemblies.

Repeat content analysis. Repeat content was identified using RepeatScout[26] followed by RepeatMasker using both nematode repeats from RepBase v17.06 (ref. 27) and the output from RepeatScout. Only hits with a Smith-Waterman score >250 were maintained. Additional repeats were then identified on the basis of abnormally high read coverage in the genome assemblies using genome sequence scanning with hysteresis triggering. Positions with read depth 20 times the mode of the read depth distribution switched the 'collapsed reads' state to on during the scanning process, and positions with read depth lower than 10 times the mode switched the 'collapsed reads' state to off. Only regions longer than 100 nucleotides were reported. Read mapping was performed by runMapping application of the Newbler suite[28]. The output was converted to SAM file format by the seq.Newbler2SAM option of the GLU package. Only the best alignment of each read was kept. Read depth was calculated by the genomeCoverageBed program of BEDTools suite[29].

RNA-Seq. RNA was prepared from one million *L. loa* microfilariae purified from the blood of a patient. Under liquid nitrogen, the microfilariae were disrupted by a stainless steel piston apparatus. Total RNA was extracted using the RNeasy Kit (Qiagen, Valencia, CA, USA). A non–strand specific complementary (cDNA) library for Illumina paired-end sequencing was prepared from ~37 ng of total RNA as previously described[30] with the following modifications. RNA was treated with Turbo DNase (Ambion, TX) and fragmented by heating at 80 °C for 3 min in 1× fragmentation buffer (Affymetrix, CA) before cDNA synthesis. Sequencing adaptor ligation was performed using 4,000 units of T4 DNA ligase (New England Biolabs, MA) at 16 °C overnight. After adaptor ligation, the resulting library was cleaned, size selected twice using 0.7× volumes of Ampure beads (Beckman Coulter Genomics, MA), enriched using 18 cycles of PCR and cleaned using 0.7× volumes of Ampure beads (Beckman Coulter Genomics, MA). The resulting Illumina sequencing library was sequenced with 76 base paired-end reads on an Illumina GAII instrument (v1.8 analysis pipeline) following the manufacturer's recommendations (Illumina, CA).

Identification of transfers (nuwts). An initial search of the *Wolbachia* of *B. malayi* genome against the *L. loa* genome was done using BLASTN with a cutoff of 1×10^{-5}. After this assembly-based search, nuclear *Wolbachia* transfers (nuwts) were identified through a screen of the *L. loa* sequencing reads as being >80% identical to *Wolbachia* sequences over 50% of the read. Searches were refined to examine reads with >50 bp match to *Wolbachia* and were manually curated to remove spurious matches that had a nematode ancestry. Reads matching the bacterial ribosomal RNA (rRNA) were removed, as they could arise from any bacterial genome that might be contaminating the sample. Regions of homology <50 bp were included if they were detected through analysis of an adjacent region with homology over >50 bp. All of the reads containing nuwts were mapped back to the *L. loa* genome to identify the consensus sequence, and the relationship was confirmed using BLASTN to NT. Phylogenetic analysis was conducted on nucleotide sequences of predicted nuwts using RAxML[30].

Annotation. Genes for both *L. loa* and *W. bancrofti* were predicted using a combination of *ab initio* gene prediction tools as previously described[31]. We also used TBLASTN to search the genome assembly against protein sequences of the following species: *C. elegans*, *C. briggsae*, *Schistosoma mansoni*, *Schistosoma japonicum* and *B. malayi* (downloaded from GenBank on February 16, 2010). The top BLAST hits are used to construct GeneWise[32] gene models. In addition, we generated gene models using available EST data from *L. Loa*, *W. bancrofti*, *O. volvulus* and *B. malayi* (downloaded from GenBank on December 2, 2009). All of these models were used as input into EVM[33] to generate combined gene predictions. To incorporate the *L. loa* RNA-Seq data, we aligned all RNA-Seq reads to the *L. loa* genome using BLAT[34]. Next we use the Inchworm module of the Trinity package[35] with default settings in genome-guided mode to assemble the reads into EST-like transcripts. These transcripts, along with the models from EVM into PASA[33], were used for gene model improvement. Gene sets were subsequently filtered to remove repeats, including genes overlapping rRNA, transfer RNA (tRNA) or output from RepeatScout[26] or TransposonPSI. Every annotated gene was given a locus identification of the form LOAG_##### (*L. loa*) or WUBG_##### (*W. bancrofti*). Pfam domains within each gene were identified using Hmmer3 (ref. 36), and gene ontology terms were assigned using BLAST2GO[37]. Secretion signals and transmembrane domains were identified using SignalP 4.0 (ref. 38) and TmHmm[39], respectively. Core eukaryotic genes were identified using CEGMA[40].

Identification of fragmented genes. Fragmented *W. bancrofti* genes were associated to their putative intact orthologs in *L. loa* or *B. malayi* by unidirectional BLAST of *W. bancrofti* peptides against peptides from the reference genome (*L. loa* or *B. malayi*). *W. bancrofti* proteins with <80% similarity to the reference, on the basis of query length, and an E value $>1 \times 10^{-10}$ were disregarded. A gene was considered fragmented if its length in *W. bancrofti* was at least 50% shorter than its length in its respective ortholog. The number of reference genome orthologs with multiple assigned fragments in *W. bancrofti* was then used to extrapolate a corrected gene count for *W. bancrofti*. An identical analysis was done for *L. loa* genes by comparison to *B. malayi*.

Synteny analysis. Whole-genome alignments of *C. elegans*, *B. malayi* and *W. bancrofti* against *L. loa* were performed by progressive Mauve[41] with default parameters. The extent of the alignment between a pair of sequences was defined as the length spanning all their respective colinear blocks. For each comparison, chromosomes or scaffolds having the longest alignment against *L. loa* scaffold number 4 (100 scaffolds from *W. bancrofti* and 30 scaffolds from *B. malayi* and *C. elegans* chromosome 3) were selected for visualization. For the systematic evaluation of synteny, pairwise syntenic blocks between the genomes of *L. loa*, *C. elegans*, *B. malayi* and *W. bancrofti* were defined by DAGchainer[42] with the minimum number of colinear genes set to three.

Gene clustering and phylogenetic analysis. We built a comparative set of genomes including those sequenced in this study and those of *P. pacificus* (from www.pristionchus.org), *C. elegans* (release 224 from WormBase), *C. briggsae* (CAAC00000000), *M. hapla* (from www.pngg.org), *B. malayi* (release 230 from WormBase), *A. suum* (published release from WormBase) and *T. spiralis* (ABIR00000000). Genes were clustered using OrthoMCL with a Markov inflation index of 1.5 and a maximum E value of 1×10^{-5} (ref. 43). Amino acid sequences of orthologs present as single copies in all genomes were aligned using MUSCLE[44] and concatenated. We then estimated

phylogenies from this data set using three methods. Parsimony bootstrapping analysis was conducted with PAUP[45] using unweighted characters and 1,000 bootstrap replicates. For maximum likelihood analysis, we first selected the JG model[46] using ModelGenerator[47] and then used the PROTCATJG model in RAxML[30] with 1,000 bootstrap replicates. For Bayesian analysis, we used MrBayes[48] with a mixed amino acid model and gamma-distributed rates. We ran the analysis with one chain for 1 million generations, sampling every 500 generations and discarding the first 25% of samples as burn in. Enrichment analyses were conducted using Fisher's exact test, and multiple comparisons were corrected using the false discovery rate[49].

Kinase classification. Initial sets of protein kinases were identified by orthology with annotated *C. elegans* kinases. Kinases without orthologs were identified in a search of the proteome against a protein kinase hidden Markov model derived from an alignment of *Dictyostelium* protein kinases[50] using a cutoff score of –66. Low-scoring sequences were additionally screened for conservation of known protein kinase sequence motifs. All protein kinases were classified using a controlled vocabulary[51,52], and classifications of filarial kinases with *C. elegans* orthologs were mapped from the curated set from the KinBase database. Kinases without orthologs in *C. elegans* were searched against the curated set using BLAST and classified if the top three hits agreed. Orthology across all nematodes was then used to identify potentially missed kinases and ensure consistent classification.

Metabolic reconstruction. In addition to the nine nematode genomes listed above, we used three additional *Wolbachia* genomes from *B. malayi* (AE017321), *D. melanogaster* (AE017196) and *C. pipiens* (AM999887). Metabolic pathways were characterized using Pathway Tools[53]. Metabolic reconstruction was performed using EFICAz2 (ref. 54) to assign Enzyme Commission numbers for each enzyme. Enzyme Commission numbers and gene names were used as input to the Pathologic software[55] with transport-identification-parser and pathway-hole-filler options set to assign MetaCyc[56] pathways for each organism. The full set of metabolic pathways for each genome is available at the WormCyc database.

25. Lennon, N.J. *et al.* A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. *Genome Biol.* **11**, R15 (2010).
26. Price, A.L., Jones, N.C. & Pevzner, P.A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21** (suppl. 1), i351–i358 (2005).
27. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
28. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
29. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
30. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
31. Haas, B.J., Zeng, Q., Pearson, M.D., Cuomo, C.A. & Wortman, J.R. Approaches to fungal genome annotation. *Mycology* **2**, 118–141 (2011).
32. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
33. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
34. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
35. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
36. Eddy, S.R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
37. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
38. Petersen, T.N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011).
39. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
40. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
41. Darling, A.E., Mau, B. & Perna, N.T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).
42. Haas, B.J., Delcher, A.L., Wortman, J.R. & Salzberg, S.L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20**, 3643–3646 (2004).
43. Li, L., Stoeckert, C.J. Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
44. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
45. Swofford, D.L. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*, version 4.0b10. (Sinauer Associates, Sunderland, Massachusetts, 2003).
46. Le, S.Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
47. Keane, T.M., Creevey, C.J., Pentony, M.M., Naughton, T.J. & McInerney, J.O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* **6**, 29 (2006).
48. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
49. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
50. Goldberg, J.M. *et al.* The dictyostelium kinome—analysis of the protein kinases from a simple model organism. *PLoS Genet.* **2**, e38 (2006).
51. Hanks, S.K. & Hunter, T. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J.* **9**, 576–596 (1995).
52. Manning, G., Whyte, D.B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912–1934 (2002).
53. Karp, P.D. *et al.* Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.* **11**, 40–79 (2010).
54. Arakaki, A.K., Huang, Y. & Skolnick, J. EFICAz2: enzyme function inference by a combined approach enhanced by machine learning. *BMC Bioinformatics* **10**, 107 (2009).
55. Karp, P.D. *et al.* Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33**, 6083–6089 (2005).
56. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **40**, D742–D753 (2012).