LETTER

Ascaris suum draft genome

Aaron R. Jex[1]*, Shiping Liu[2]*, Bo Li[2]*, Neil D. Young[1]*, Ross S. Hall[1], Yingrui Li[2], Linfeng Yang[2], Na Zeng[2], Xun Xu[2], Zijun Xiong[2], Fangyuan Chen[2], Xuan Wu[2], Guojie Zhang[2], Xiaodong Fang[2], Yi Kang[2], Garry A. Anderson[1], Todd W. Harris[3], Bronwyn E. Campbell[1], Johnny Vlaminck[4], Tao Wang[4], Cinzia Cantacessi[1], Erich M. Schwarz[5], Shoba Ranganathan[6], Peter Geldhof[4], Peter Nejsum[7], Paul W. Sternberg[5], Huanming Yang[2], Jun Wang[2], Jian Wang[2] & Robin B. Gasser[1]

Parasitic diseases have a devastating, long-term impact on human health, welfare and food production worldwide. More than two billion people are infected with geohelminths, including the roundworms *Ascaris* (common roundworm), *Necator* and *Ancylostoma* (hookworms), and *Trichuris* (whipworm), mainly in developing or impoverished nations of Asia, Africa and Latin America[1]. In humans, the diseases caused by these parasites result in about 135,000 deaths annually, with a global burden comparable with that of malaria or tuberculosis in disability-adjusted life years[1]. *Ascaris* alone infects around 1.2 billion people and, in children, causes nutritional deficiency, impaired physical and cognitive development and, in severe cases, death[2]. *Ascaris* also causes major production losses in pigs owing to reduced growth, failure to thrive and mortality[2]. The *Ascaris*–swine model makes it possible to study the parasite, its relationship with the host, and ascariasis at the molecular level. To enable such molecular studies, we report the 273 megabase draft genome of *Ascaris suum* and compare it with other nematode genomes. This genome has low repeat content (4.4%) and encodes about 18,500 protein-coding genes. Notably, the *A. suum* secretome (about 750 molecules) is rich in peptidases linked to the penetration and degradation of host tissues, and an assemblage of molecules likely to modulate or evade host immune responses. This genome provides a comprehensive resource to the scientific community and underpins the development of new and urgently needed interventions (drugs, vaccines and diagnostic tests) against ascariasis and other nematodiases.

We sequenced the *A. suum* genome at ~80-fold coverage (Supplementary Fig. 1), producing a final draft assembly of 272,782,664 base pairs (bp) (N50 = 407 kilobases, kb; N90 = 80 kb; 1,618 contigs of >2 kb) (Table 1) with a mean GC-content of 37.9%. This genome has few repetitive sequences (about 4.4% of the total assembly) relative to that reported for other metazoan genomes sequenced to date[3–6], probably as a result of chromatin diminution[7]. We identified 424 distinct retrotransposon sequences (see Supplementary Tables 1–3) representing at least 22 families (8 long terminal repeats (LTRs), 12 long interspersed elements (LINEs) and 2 short interspersed elements (SINEs)), with *Gypsy*, *Pao* and *Copia* classes predominating for LTRs ($n = 97$, 85 and 60, respectively) and CR1, L1, and reverse transcriptase encoding RTE-RTE classes predominating for non-LTRs ($n = 29$, 28 and 21, respectively). We also identified eight families of DNA transposons (91 distinct sequences in total), of which *MuDr*, *En-Spm* and *Merlin* ($n = 12$, 9 and 8, respectively) predominated. We predicted 18,542 genes (14,783 supported by transcriptomic data), with a mean total length of 6.5 kb, exon length of 153 bp and a mean of 6.4 exons per gene (see Supplementary Fig. 2). Compared with the nematodes (roundworms) *Caenorhabditis elegans*[3], *Pristionchus pacificus*[8], *Brugia malayi*[9] or *Meloidogyne hapla*[10], overall, the *A. suum* genes are significantly longer (see Supplementary Table 2), relating primarily to expansions of intronic regions (mean 1.1 kb).

Most (78.2%) of the predicted *A. suum* genes (Fig. 1) have a homologue (BLASTp cut-off $\leq 10^{-5}$) either in *C. elegans* ($n = 12,779$; 68.9%), *B. malayi* (12,853; 69.3%), *M. hapla* (10,482; 56.5%) or *P. pacificus* (11,865; 64.0%), with 8,967 being homologous among all species examined, and 4,042 (21.8%) being 'unique' to *A. suum* (see Fig. 1). Of the genes with homology to *C. elegans* or *B. malayi*, ~50%

Table 1 | Features of the *Ascaris suum* draft genome

Estimated genome size in megabases	309
Total number of base pairs within assembled scaffolds	272,782,664
N50 length in bp; total number >2 kb in length	407,899; 1,618
N90 length in bp; total number >N90 length	80,017; 748
GC content of whole genome (%)	37.9
Repetitive sequences (%)	4.4
Proportion of genome that is coding (exonic; including introns) (%)	5.9; 44.2
Number of putative coding genes	18,542
Gene size (mean bp)	6,536
Average coding domain length (mean bp)	983
Average exon number per gene (mean)	6
Gene exon length (mean bp)	153
Gene intron length (mean bp)	1,081
GC content in coding regions (%)	45
Number of transfer RNAs	255

N50 means 50% of all nucleotides in the assembly are within contigs of ≥408 kb. N90 means 90% of all nucleotides in the assembly are within contigs of ≥80 kb. Genome size estimated on the basis of *k*-mer (see online-only Methods) frequency.
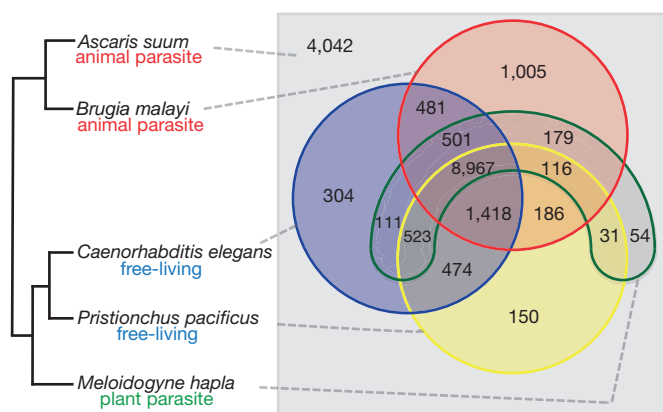

Figure 1 | Venn diagram summarizing the overlapping homology between the *Ascaris suum* gene set and those of other nematodes. Grey box (right) represents genes unique to *A. suum*, relative to *Brugia malayi* (red circle), *Caenorhabditis elegans* (blue circle), *Meloidogyne hapla* (green arc) and/or *Pristionchus pacificus* (yellow circle). The phylogram (left) displays the evolutionary relationships currently proposed among the nematodes.

[1]Faculty of Veterinary Science, The University of Melbourne, Parkville, Victoria 3010, Australia. [2]BGI-Shenzhen, Shenzhen, 518083, China. [3]Ontario Institute for Cancer Research, MaRS Centre, South Tower, 101 College Street, Suite 800, Toronto, Ontario, M5G 0A3, Canada. [4]Laboratory of Parasitology and Parasitic Diseases, University of Ghent, Merelbeke B-9820, Belgium. [5]California Institute of Technology, Pasadena, California, 91125, USA. [6]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, New South Wales 2109, Australia. [7]Faculty of Life Sciences, University of Copenhagen, Copenhagen DK-2200, Denmark.
*These authors contributed equally to this work.

and 44%, respectively, were determined to represent one-to-one orthologues[11] (see Supplementary Data 1). For these orthologues (on scaffolds exceeding one megabase, 1 Mb, in size), we explored synteny for *A. suum* and *B. malayi* by pairwise comparison with *C. elegans* (see Supplementary Data 1). The findings show that interchromosomal gene rearrangements in *A. suum* are relatively rare and occurred less frequently in *A. suum* than in *B. malayi*[9] relative to *C. elegans* since their evolutionary divergence[12]. In contrast, intrachromosomal rearrangements were relatively common and comparable in frequency to those inferred for *B. malayi*[9]. Overall synteny was significantly higher between *A. suum* and *B. malayi* (~15%) than between either species and *C. elegans* (~3%), which is consistent with current knowledge of the evolutionary relationships among these three species[12]. Interestingly, of these *C. elegans* orthologous genes, 532 and 483 were exclusive to the current assemblies of the *A. suum* and *B. malayi* genomes, respectively (Supplementary Data 2). Although there were no homology matches between these two exclusive subsets of orthologues, they shared striking similarity in functional ontology (biological process), being linked predominantly to growth, reproduction, development and/or morphogenesis. There is clear evidence of plasticity in the germline of metazoans[13], with cases of products from non-homologous genes in different species having analogous function(s). Therefore, we hypothesize that these two unique gene subsets relate to differences in reproductive biology (oviparity versus viviparity) and life history (direct versus indirect) between *A. suum* and *B. malayi*. Clearly, this proposal warrants testing and functional validation in *C. elegans* and/or in *Ascaris*.

Of the entire *A. suum* gene set, 2,370 genes had an orthologue (BLASTp cut-off $\leq 10^{-5}$) belonging to one of 279 known biological (KEGG; see online-only Methods) pathways (Supplementary Data 3). Mapping to pathways in *C. elegans* indicated a full complement of molecules; by inference, the vast majority (95%) of the *A. suum* euchromatin is represented in the present genomic assembly, an inference that is supported by our transcriptomic data (Supplementary Tables 4 and 5). We were able to assign possible functions (such as for enzymes, receptors, channels and transporters; Supplementary Fig. 3, Supplementary Table 6 and Supplementary Data 4) to 13,503 (72.8%) of the genes predicted for *A. suum* (Fig. 2). For these genes, we predicted 456 peptidases belonging to five major classes (aspartic, cysteine, metallo-, serine and threonine), with the metallo- ($n = 184$:
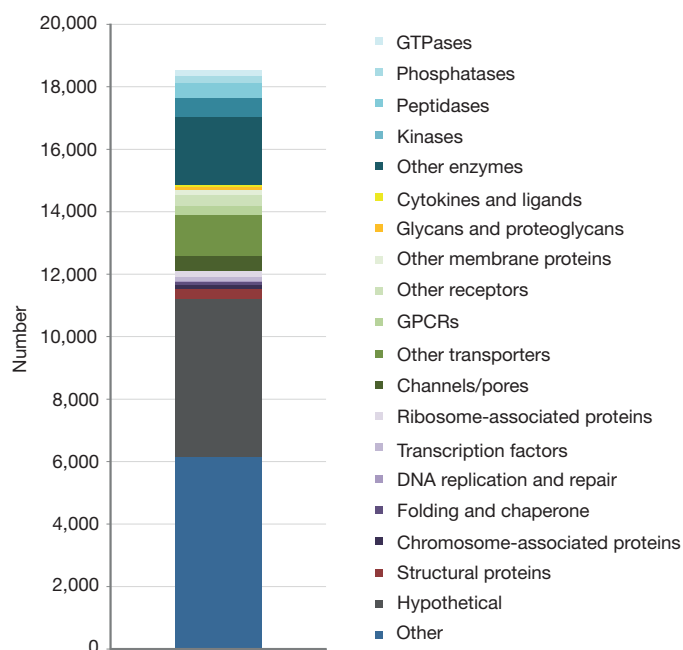
41.0%) and serine proteases ($n = 132$: 30.0%) predominating (Supplementary Data 4). Notably, the secreted peptidases (such as the M12 'astacins', the S9 and S33 serine proteases, and the C1 and C2 cysteine proteases) are abundantly represented, and have key roles in tissue invasion and degradation (for example, during migration and/or feeding) and/or immune evasion/modulation in many parasites[14,15].

In addition, we identified 609 kinases and 257 phosphatases, respectively (Supplementary Data 4). All major classes of kinases are represented, with the tyrosine (TK: $n = 94$), casein (CK1: $n = 83$), CMGC ($n = 67$) and CAMK ($n = 54$) being most abundant in *A. suum*. The phosphatome includes 17 receptor and 68 conventional tyrosine, 64 serine/threonine and 39 dual-specificity phosphatases. On the basis of homology with molecules in *C. elegans*, 169 GTPases are encoded in the *A. suum* genome, including 135 small GTPases (Ras superfamily) representing the Rab ($n = 36$), Ras ($n = 35$; plus 8 Ras-like), Rho ($n = 17$; plus 9 Rho-like) or Ran ($n = 6$) subfamilies. Examples of these homologues include *eft-1*, *fzo-1*, *glo-1* and *rho-1*, which have essential roles in embryonic, larval and/or reproductive development (see www.wormbase.org).

Given their key roles, many of these enzymes are proposed as targets for anti-parasitic compounds and/or vaccines[16–18]. Equally, the range of receptor and channel proteins identified here are interesting because many common anthelmintics bind such targets[19]. Here, we predicted 279 G protein-coupled receptors (GPCRs) for *A. suum* and 477 channel or pore proteins (Supplementary Data 4), including 272 voltage-gated and 98 ligand-gated ion channels. Many voltage-gated ion channels are known targets for nematocidal drugs, such as macrocyclic lactones (for example, ivermectin) and levamisole, and an aminoacetonitrile derivative, monepantel, is the most recent example of a highly effective nematocide that binds to a ligand-gated ion channel[19]. Importantly, in the *A. suum* gene set, we found a homologue (*acr-23*) of the *C. elegans* monepantel receptor[19], suggesting that this drug may kill *A. suum*. In addition, we detected 462 transporters (for example, small molecule porter proteins), of which the major facilitator ($n = 155$), cation symporter ($n = 71$) and resistance-nodulation-cell division ($n = 56$) superfamilies were most abundant (Supplementary Data 4).

Excretory/secretory (E/S) peptides are central to understanding parasite–host interactions. We predicted the secretome of *A. suum* to comprise 775 proteins with diverse functions (Supplementary Data 5). Notable among them are 68 secreted proteases, including 20 SC clan serine proteases (S9 and S33 families), 18 MA clan metallo-proteases (M10, M12 and M41 families) and 5 CA/CD clan cysteine proteases (C1 and C13 families); see http://merops.sanger.ac.uk/ for clan definitions.

Secreted proteases have known roles in host-tissue degradation, required for feeding, tissue-penetration and/or larval migration for a range of helminths[14], including *Ascaris*[2]. In addition, they are involved in inducing and modulating host immune responses against parasitic helminths[15], which are often Th2-biased[20]. From the current understanding of these responses[15], we compiled a comprehensive list of *A. suum* E/S proteins homologous to helminth-secreted peptides with important immunogenic or immunomodulatory roles in host animals (Supplementary Table 7 and Supplementary Data 6). Such homologues represent about half of the predicted *A. suum* secretome. Most abundant among them are O-linked glycosylated proteins ($n = 300$), many of which are heavily targeted by immunoglobulin (Ig) M antibodies and bound by various pattern recognition receptors associated with host dendritic cells responsible for the induction of a Th2 immune response[15].

Other members of the *A. suum* secretome are predicted to direct or evade immune responses. These peptides include a close homologue of the E/S-62 leucyl aminopeptidase of the filarioid nematode *Acanthocheilonema viteae*, which has been shown to inhibit B-cell, T-cell and mast cell proliferation/responses, promote an alternative activation of the host macrophages, through the inhibition of the Toll-like receptor signalling pathway, and induce a Th2 response through

Figure 2 | The major protein classes representing the *Ascaris suum* gene set.

Legend:
- GTPases
- Phosphatases
- Peptidases
- Kinases
- Other enzymes
- Cytokines and ligands
- Glycans and proteoglycans
- Other membrane proteins
- Other receptors
- GPCRs
- Other transporters
- Channels/pores
- Ribosome-associated proteins
- Transcription factors
- DNA replication and repair
- Folding and chaperone
- Chromosome-associated proteins
- Structural proteins
- Hypothetical
- Other

the inhibition of IL-12p70 production by dendritic cells[15]. Additional, immunomodulatory molecules predicted for *A. suum* (BLASTp cut-off $\leq 10^{-5}$) include homologues of another B-cell inhibitor (that is, the *B. malayi* cystatin CPI-2), several TGF-β and macrophage initiation factor mimics, numerous neutrophil inhibitors, various oxidoreductases, and five close homologues of platelet anti-inflammatory factor α (ref. 15). Some *A. suum* E/S peptides are predicted to be involved in immune evasion; for instance, some mask parasite antigens by mimicking host molecules (such as several C-type lectins with close homology to vertebrate macrophage mannose or CD23 (low affinity IgE receptors[15]).

Taken together, these data indicate that *A. suum* has a large arsenal of E/S proteins that are likely to be involved directly in manipulating, blocking and/or evading immune responses in the host. Understanding the immunomolecular interplay between *A. suum* and its host, early in infection, particularly during hepatopulmonary migration, should pave the way for designing prophylactic interventions, such as vaccination.

Ascaris larvae undertake an extensive migration through their host's body before they establish as adults in the small intestine. Following the ingestion of infective eggs and their gastric passage, third-stage larvae (L3s)[21] hatch from eggs in the gut and penetrate the intestinal wall; they then undergo, via the bloodstream, an arduous hepatopulmonary migration. The complexity of this migration coincides with important developmental changes in the nematode[2]. Clearly, this migration requires tightly regulated transcriptional changes in the parasite. We explored this aspect by characterizing the transcription profiles of infective L3s (from eggs), L3s from the liver or lungs of the host, and fourth-stage larvae (L4s) from the small intestine (Supplementary Fig. 4, Supplementary Data 7). Notable among genes enriched during larval migration are various secreted peptidases linked to tissue-penetration and degradation during feeding and/or migration[14], including three C1/C2, five M1, eight M12, fourteen S9 and five S33 clan members. Considering the complex nature of larval migration, a key role for molecules associated with chemosensory pathways is highly likely. Such molecules have been studied extensively in *C. elegans*[22], with numerous homologues being identified here in larval transcripts (Supplementary Data 7). With few exceptions, all of these homologues relate to olfactory chemosensation of volatile compounds (for example, alcohols, aldehydes or ketones), suggesting that the olfactory detection of molecular gradients is central to the navigation of *A. suum* larvae during migration. Lastly, considering the substantial host attack against migrating *Ascaris* larvae, E/S proteins probably play crucial roles in immune modulation and/or evasion during hepatopulmonary migration. Many such genes, including *Bm-alt-1*, *Bm-cpi-2* and *mif-4*, are highly transcribed in *A. suum* larvae (see Supplementary Data 7), particularly in migrating L3s.

Because of the large size of the adult nematode (10–15 cm), we were able to explore transcription in the musculature and reproductive tracts of adult male and female *A. suum* individuals as well (Supplementary Fig. 5 and Supplementary Data 8). Among the male-enriched transcripts is a range of genes associated specifically with sperm and/or spermatogenesis, including *fer-1*, *spe-4*, *spe-6*, *spe-9*,

spe-10, *spe-15* and *spe-41*, *alg-4* and *msp-57* (see www.wormbase.org). Notable among the female-enriched transcripts is a large variety of genes associated with oogenesis/egg-laying (such as *cat-1*, *unc-54*, *cbd-1* and *pqn-74*), vulval development (such as *noah-1*, *nhr-25*, *cog-1* and *pax-3*) and/or embryogenesis (such as *cam-1* and *unc-6*; see www.wormbase.org). Although the functions of these genes have been explored in *C. elegans* (primarily a hermaphroditic nematode), this detailed insight into the tissue-specific transcription for a dioecious nematode is a major advance.

Analyses of these RNA-seq data revealed 163,777 single nucleotide polymorphisms (SNPs) in coding regions of the *A. suum* genome; 61% of them were synonymous, 7% non-synonymous and <0.1% termination codons (Supplementary Data 9). Some of the most variable genes in *A. suum* encoded ribosomal proteins ($n = 44$), translation initiation factor (*tif*) eIF-3 subunits 3 and 5, *tif* TFIIH subunit H2 and *tif* IF-2, galectin-4 and galectin-9, the latter two of which are probably linked to immune evasion[15] and may indicate that antigenic variation is among the many strategies apparent in *Ascaris* to combat the host immune response. Interestingly, the high nucleotide variability linked to the key elements of translation machinery did not relate to a bias in synonymous SNPs, suggesting that many mutations accumulate in particular 'hotspots' and/or are tolerated, but do not compromise either the structure or the function of this machinery. The least variable genes encoded various (druggable)[16,17] serine/threonine phosphatases ($n = 17$) as well as numerous receptors, channels and transporters, for which there was an unusually strong bias towards synonymous SNPs, reinforcing their potential as intervention targets.

Given our present reliance on a small number of drugs (for example, piperazine, pyrantel, albendazole and mebendazole) for the treatment of ascariasis, their repeated or excessive use might lead to resistance in *Ascaris* populations to some or all of these compounds[23]. As few new anthelmintics (that is, aminoacetylnitriles[19] and cycloocto-depsipeptides[24]) have been discovered in the past two decades using traditional screening methods, an effective, alternative means of drug discovery is urgently needed[23]. Genome-guided drug target or drug discovery has major potential to complement conventional screening and re-purposing. The goal of genome-guided analysis is to identify genes or molecules whose inactivation by one or more drugs will selectively kill parasites but not harm their host.

Because most parasitic nematodes are difficult to produce or maintain outside of their host, or to subject to gene-specific silencing by RNAi[23] or morpholinos[25,26], direct functional assessment of essentiality (that is, they are needed for the nematode's survival) is not yet practical. However, essentiality can be inferred from functional information for model organisms (for example, lethality in *C. elegans* and *D. melanogaster*)[27], and this approach has indeed yielded effective targets for nematocides[16]. In *Ascaris*, we identified 629 proteins (Supplementary Data 10) with essential homologues in *C. elegans* and *D. melanogaster* (linked to lethal phenotypes upon gene perturbation). Among these are 87 channels or transporters (including 44 voltage-gated ion channels), which represent protein classes most successfully targeted for anthelmintic compounds, including macrocyclic lactones, levamisoles and aminoacetonitrile

Table 2 | Druggable candidates in the *Ascaris suum* draft genome

Protein or chokepoint	Subtype (number of molecules)	Total number
GTPase	Small GTPase (22); Ras (13); Rab (5); Rho (3); Ras-like (1); others (2)	46
Kinase	TK (8); AGC (3); CAMK (2); TKL (2); STE (1); other (1)	17
Peptidase	A22A (5); M14B (3); M12B (2); M67A (1); C14A (1); C50 (1); M12A (1); M13 (1); T01A (1); C46 (1); S33 (1)	19
Phosphatase	STP (28); cPTP (4); DSP (3)	35
Transporters and channels	Channels and pores (30); primary active transporters (24); incompletely characterized transport system proteins (22); accessory factors involved in transport (5); electrochemical potential-driven transporters (5); group translocators (1)	87
'Lethal' chokepoints	CDP-diacylglycerol-inositol 3-phosphatidyltransferase	1
	G protein-coupled receptor kinase 5	1
	Phosphoribosylformylglycinamidine synthase	1
	Inosine-5′-monophosphate dehydrogenase	1
	Phospho-*N*-acetylmuramoyl pentapeptide transferase	1

Candidates were inferred from essentiality prediction and metabolic chokepoint analysis.

derivatives[19,28]. Also notable are 46 GTPases, 35 phosphatases (including PP1 and PP2A homologues, as targets for norcantharidin analogues)[16], 17 kinases and 19 peptidases (Table 2).

In addition to essentiality-based prediction, an alternative strategy has been to infer enzymatic chokepoints intrinsic to the complete metabolome of a parasite[29]. Such chokepoints are defined as enzymatic reactions that uniquely produce and/or consume a molecular compound, using the strategy that the disruption of such enzymes would lead to the toxic build-up (that is, for unique substrates) or starvation (that is, for unique products) of metabolites within cells. Pathway analysis identified 225 likely chokepoints linked to genes predicted to be essential in *A. suum* (Supplementary Data 10). We gave the highest priority to targets predicted from single-copy genes in the *A. suum* genome, reasoning that lower allelic variability would exist within populations and would thus be less likely to give rise to drug resistance.

Using this strategy, we identified five high-priority drug targets for *A. suum* (see Table 2 and Supplementary Data 10) that, given their conservation with *C. elegans* and *D. melanogaster*, are likely to be relevant in relation to many other parasitic helminths. Conspicuous among them is IMP dehydrogenase (GMP reductase), which has a variety of inhibitors (for example, mycophenolic acid analogues[30]) that could be tested for ascaricidal effects. Clearly, the druggable genome of *Ascaris* now provides a solid basis for rational drug design, aimed at controlling parasitic nematodes of major socioeconomic impact worldwide.

In conclusion, we have characterized the genome of *A. suum*, a major parasite of one of the world's most important food animals (pig) and the closest relative of *A. lumbricoides*, which infects about 1.2 billion people globally[1,2]. Intriguingly, the present *A. suum* draft genome exhibits unusually low repeat content and lacks Tas2 transposons[7]. These characteristics probably relate to the chromatin diminution described previously for some ascaridoids[7], indicating that our assembly represents the somatic genome of this parasite. The precise mechanism governing this diminution is not yet understood. Although the chromatin lost during this process is not fully characterized, there appears to be a significant loss in repeat content[7], consistent with the present assembly. Notably, the present gene set inferred for *A. suum* includes *fert-1* and *rpS19G*, which, although originally proposed to be germline-specific[7], were transcribed in all adult libraries sequenced here. This finding suggests that the genomic content lost during diminution might vary among individuals or tissues, and is a stimulus to investigate chromatin diminution between and among individual cells (that is, sperm or eggs), stages and tissue types of *A. suum*. Importantly, the present study, showing that a high-quality genomic assembly can be achieved using an approach based on whole-genome amplification, provides unique prospects for exploring diminution in detail, using the present genome as a reference.

In addition, our sequencing effort has characterized a broad range of key classes of molecules of major relevance to understanding the molecular biology of *A. suum* and the exquisite complexities of the host–parasite interplay on an immunobiological level. This work paves the way for future fundamental molecular explorations and the design of new methods for the treatment and control of one of the world's most important parasitic nematodes. This focus is now crucial, given the major impact of *Ascaris* and other soil-transmitted helminths, which affect billions of people and animals worldwide. Although these parasites are seriously neglected, genomic and post-genomic approaches provide new hope for the discovery of intervention strategies, with major implications for improving global health.

METHODS SUMMARY

We sequenced the genome of *A. suum* using Illumina technology from genomic DNA from the reproductive tract of a single adult female. From six paired-end sequencing libraries (insert sizes: 0.17 kb to 10 kb; see Supplementary Tables 1 and 2), we generated 39 Gb of useable short-read sequence data, equating to ~80-fold coverage of the 273-Mb genome. We assembled the short reads, constructed

scaffolds in a step-by-step manner, and then closed intra-scaffold gaps[5]. Transposable elements, non-coding RNAs and the protein-coding gene set were inferred using a combination of predictive modelling and a homology-based approach. Orthology and synteny analyses were conducted using established methods[9,11]. We sequenced messenger RNA from infective L3s (from eggs), migrating L3s from the liver or lungs of the host, and L4s from the small intestine, as well as muscle and reproductive tissues from adult male and female *A. suum*, and used these data to aid gene predictions, define SNPs and explore key molecules associated with larval migration, reproduction and development. All proteins predicted from the gene set were annotated using databases for conserved protein domains, gene ontology annotations and model organisms (that is, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus*). Essentiality and drug target predictions were conducted using established or in-house methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Hotez, P. J., Fenwick, A., Savioli, L. & Molyneux, D. H. Rescuing the bottom billion through control of neglected tropical diseases. *Lancet* **373**, 1570–1575 (2009).
2. Crompton, D. W. *Ascaris* and ascariasis. *Adv. Parasitol.* **48**, 285–375 (2001).
3. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
4. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
5. Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
6. Mitreva, M. *et al.* The draft genome of the parasitic nematode *Trichinella spiralis*. *Nature Genet.* **43**, 228–235 (2011).
7. Müller, F. & Tobler, H. Chromatin diminution in the parasitic nematodes *Ascaris suum* and *Parascaris univalens*. *Int. J. Parasitol.* **30**, 391–399 (2000).
8. Dieterich, C. *et al.* The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nature Genet.* **40**, 1193–1198 (2008).
9. Ghedin, E. *et al.* Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317**, 1756–1760 (2007).
10. Opperman, C. H. *et al.* Sequence and genetic map of *Meloidogyne hapla*: a compact nematode genome for plant parasitism. *Proc. Natl Acad. Sci. USA* **105**, 14802–14807 (2008).
11. Kuzniar, A., van Ham, R. C. H. J., Pongor, S. & Leunissen, J. A. M. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.* **24**, 539–551 (2008).
12. Blaxter, M. L. *et al.* A molecular evolutionary framework for the phylum Nematoda. *Nature* **392**, 71–75 (1998).
13. Ewen-Campen, B., Schwager, E. E. & Extavour, C. G. The molecular machinery of germ line specification. *Mol. Reprod. Dev.* **77**, 3–18 (2010).
14. McKerrow, J. H., Caffrey, C., Kelly, B., Loke, P. & Sajid, M. Proteases in parasitic diseases. *Annu. Rev. Pathol.* **1**, 497–536 (2006).
15. Hewitson, J. P., Grainger, J. R. & Maizels, R. M. Helminth immunoregulation: the role of parasite secreted proteins in modulating host immunity. *Mol. Biochem. Parasitol.* **167**, 1–11 (2009).
16. Campbell, B. E. *et al.* Norcantharidin analogues with nematocidal activity in *Haemonchus contortus*. *Bioorg. Med. Chem. Lett.* **21**, 3277–3281 (2011).
17. 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).
18. Renslo, A. R. & McKerrow, J. H. Drug discovery and development for neglected parasitic diseases. *Nature Chem. Biol.* **2**, 701–710 (2006).
19. Kaminsky, R. *et al.* A new class of anthelmintics effective against drug-resistant nematodes. *Nature* **452**, 176–180 (2008).
20. Maizels, R. M. & Yazdanbakhsh, M. Immune regulation by helminth parasites: cellular and molecular mechanisms. *Nature Rev. Immunol.* **3**, 733–744 (2003).
21. Geenen, P. L. *et al.* The morphogenesis of *Ascaris suum* to the infective third-stage larvae within the egg. *J. Parasitol.* **85**, 616–622 (1999).
22. Bargmann, C. I. Chemosensation in *C. elegans* in *Wormbook* (ed. The *C. elegans* Research Community) (2006); http://www.wormbook.org.
23. Keiser, J. & Utzinger, J. The drugs we have and the drugs we need against major helminth infections. *Adv. Parasitol.* **73**, 197–230 (2010).
24. Harder, A. *et al.* Cyclooctadepsipeptides—an anthelmintically active class of compounds exhibiting a novel mode of action. *Int. J. Antimicrob. Agents* **22**, 318–331 (2003).
25. Heasman, J. Morpholino oligos: making sense of antisense? *Dev. Biol.* **243**, 209–214 (2002).
26. Geldhof, P. *et al.* RNA interference in parasitic helminths: current situation, potential pitfalls and future prospects. *Parasitology* **134**, 609–619 (2007).
27. Lee, I. *et al.* A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nature Genet.* **40**, 181–188 (2008).
28. 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).
29. Berriman, M. *et al.* The genome of the blood fluke *Schistosoma mansoni*. *Nature* **460**, 352–358 (2009).

30. Chen, L., Wilson, D. J., Labello, N. P., Jayaram, H. N. & Pankiewicz, K. W. Mycophenolic acid analogs with a modified metabolic profile. *Bioorg. Med. Chem.* **16,** 9340–9345 (2008).

METHODS

Sample procurement, preparation and storage. All specimens of *A. suum* were collected from pigs (*Sus scofra*) with naturally acquired infections in Victoria, Australia (adult nematodes) and Ghent, Belgium (larval stages). L3s and L4s were also collected from the liver or lung and from the small intestine, respectively, of pigs, using established procedures[31,32]. Nematodes were washed extensively in sterile physiological saline (37 °C), snap-frozen in liquid nitrogen and then stored at −70 °C until use.

DNA isolation, sequencing and quality control. Total genomic DNA was isolated from the reproductive tract of a single adult female of *A. suum* using a sodium-dodecyl sulphate/proteinase K digestion[33] followed by phenol-chloroform extraction and ethanol precipitation[34]. Total DNA yield was determined using the Qubit fluorometer double-stranded DNA HS Kit (Invitrogen). DNA integrity was verified with a 2100 Bioanalyser (Agilent). Short-insert (170 bp and 500 bp) and mate-pair (800 bp, 2 kb, 5 kb and 10 kb) genomic DNA libraries were prepared and paired-end sequenced using TruSeq chemistry on a HiSeq 2000 (Illumina). Whole-genome amplification, employing the REPLI-g Midi Kit (Qiagen), was used to produce (from 200 ng of genomic template) the required amount of DNA for the construction of the 2-kb, 5-kb and 10-kb libraries (Supplementary Fig. 6). The sequence data generated from each of the six libraries were verified, and low-quality sequences, base-calling duplicates and adapters removed. The size of the genome and the heterozygosity rate were estimated by establishing the frequency of occurrence of each 17-bp *k*-mer (a unique sequence of *k* (that is, 17) nucleotides in length) within the genomic sequence data set (from the 170-bp library) using an established method[5]. Genome size was estimated using a modification of the Lander–Waterman algorithm[35], where the haploid genome length in base pairs is $G = (N \times (L − K + 1) − B)/D$, where N is the read length sequenced in base pairs, L is the mean length of sequence reads, K is the *k*-mer length (17 bp) and B is the number of *k*-mers occurring less than four times (Supplementary Fig. 7). Heterozygosity was evaluated throughout the genome assembly by assessing the distribution of the *k*-mer frequency in the sequence data set.

RNA isolation, sequencing and assembly. We obtained total RNAs from egg-L3s ($n \approx 500,000$), liver-L3s ($n \approx 60,000$), lung-L3s ($n \approx 80,000$) or L4s ($n \approx 30,000$) and from the somatic musculature or reproductive tract of each of two adult male and two adult female *A. suum* using the TriPure reagent (Roche), and both yield and quality were verified by 2100 BioAnalyser (Agilent). Polyadenylated (polyA+) RNA was purified from 10 µg of total RNA using Sera-mag oligo(dT) beads, fragmented to a size of 300–500 bp, reverse-transcribed using random hexamers, end-repaired and adaptor-ligated, according to the manufacturer's protocol (Illumina). Ligated products of ~400 bp were excised from agarose and then PCR-amplified (15 cycles), as recommended. Products were purified over a MinElute column (Qiagen) and subjected to paired-end RNA-seq using TruSeq chemistry on a HiSeq 2000 (Illumina) and assessed for quality and adaptor sequence. Transcripts were assembled from RNA-seq data using Oases[36]. All transcripts were used to assess the completeness of the genome assembly and to predict genes.

Genomic assembly and quality control. Following sequencing, all DNA-sequence reads were corrected based on *k*-mer (=17) distribution[5]. Briefly, sequence reads were removed if >10% of bases were ambiguous (represented by the letter N) or multiple adenosine monophosphates (poly-A), and all remaining reads were filtered on the basis of Phred quality. For small insert-size libraries (that is, <800 bp), additional reads were removed from the final data set if >65% of bases were of a low Phred quality (<8). For large insert libraries (2 kb, 5 kb and 10 kb), reads were removed from the final data set if >80% of bases were of a low Phred quality (<8). Duplicate (that is, identical) reads and partial reads representing the Illumina adaptor sequence were also removed, as were reads from the 500-bp library representing paired reads found to overlap by >10 bp (allowing for a 10% mismatch). Corrected and filtered data were assembled into contigs using SOAPdenovo[5], and joined iteratively into scaffolds using a step-wise process (see Supplementary Fig. 8), using the paired reads generated from each library; local assemblies were used to close all gaps. Each nucleotide position in the final assembly was assessed for accuracy by aligning all filtered reads to the scaffolds using SOAP2aligner[37], allowing for up to five mismatches per read. The depth of coverage and repeat content were assessed initially by sliding-window analysis and presented as a frequency distribution (Supplementary Fig. 9). GC-content was estimated using 10-kb non-overlapping sliding windows, and GC-bias[38] was assessed based on a frequency distribution of these data (Supplementary Fig. 10). To assess the completeness of the genome assembly, RNA-seq data representing each of the organs (that is, musculature and reproductive tract), genders and/or stages of *A. suum* sequenced were mapped to the final assembly using the BLAST-like Alignment Tool (BLAT)[39].

Assessment of repeat content and annotation of non-coding RNA. Following genome assembly, tandem repeats were identified using the Tandem Repeats Finder program[40]. Transposable elements were predicted using a combination of homology-based comparisons (using RepeatMasker[41]) and *de novo* approaches (using LTR_FINDER[42], PILER[43] and RepeatScout[44]), with a consensus population of predicted repetitive elements, constructed in RepeatScout using fit-preferred alignment scores. Low-frequency repeats (≤25) and multi-copy genes (in the repeat element library) were filtered using RepeatMasker, producing a non-redundant sequence file, which was then used to identify and classify additional homologous repeats in the genome.

Gene prediction, and synteny and genetic variation analysis. The *A. suum* protein-coding gene set was inferred using *de novo*-, homology- and evidence-based (that is, transcriptomic) approaches. *De novo* gene prediction was performed on a repeat-masked genome using three programs (Augustus, GlimmerHMM and SNAP)[5]; training models were generated from a subset of the transcriptomic data set representing 1,355 distinct genes. Homology-based prediction was conducted by comparison with complete genomic data for *Caenorhabditis elegans*[3], *Pristionchus pacificus*[8] and *Brugia malayi*[9] using a multi-phase strategy, in which (1) all putative homologous gene sequences were preliminarily identified from alignments with protein sequences representing the complete gene set of each of the reference genomes (the longest transcripts were chosen to represent each gene) by TblastN (*e*-value cut-off: $10^{−5}$) and grouped into gene-like structures using genBlastA[45]; (2) regions representing these putative genes, and flanking regions (3,000 bp) at the 5′- and 3′-ends of each predicted gene, were extracted from the assembly and aligned to the 'parent' sequences derived from the reference genomes using Genewise[46]; (3) all single-exon genes predicted to have arisen from a retrotransposition and containing at least one frame-shift error or representing incomplete coding domains of <150 bp as well as all multi-exon genes containing more than two frame-shift errors and/or representing incomplete coding domains of <100 bp, were discarded. Evidence-based gene prediction was conducted by aligning all RNA-seq data generated herein against the assembled genome using TopHat[47], with cDNAs predicted from the resultant data using Cufflinks[48]. Following the prediction of genes, a non-redundant gene set representing homology-based, *de-novo*-predicted and RNA-seq-supported genes, was generated using Glean (http://sourceforge.net/projects/glean-gene)[5]. All Glean-predicted genes were retained, as were all genes supported by RNA-seq data and those predicted using two or more *de novo* methods (that is, Augustus, GlimmerHMM and/or SNAP). The open reading frame of each gene was predicted using BestORF (www.softberry.com). To assess the quality and accuracy of the predicted gene set, we examined the length-distribution of all genes, coding sequences, exons and introns, and the distribution of exon numbers for individual genes, and then compared these parameters with those calculated for the published gene sets of *B. malayi*, *C. elegans*, *P. pristionchus* and *M. incognita* (Supplementary Fig. 4).

Following prediction of the finalized gene set, we conducted pairwise analysis of the overall synteny existing between/among the large (>1 Mb) assembly scaffolds for *B. malayi* and *A. suum* relative to the complete *C. elegans* chromosomes. This analysis was undertaken by conducting pairwise alignments among all *A. suum* or *B. malayi* (WS220 assembly: ftp://ftp.sanger.ac.uk/pub2/wormbase/releases/WS220/genomes/b_malayi/) scaffolds larger than 1 Mb in size and the *C. elegans* chromosomes using LASTz (http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html), which were then joined using CHAINNET[49] and output as a .axt alignment from which large-syntenic regions were defined. The resulting alignment files were used to construct synteny images on scaleable vector graphics format using customized perl scripts (ZX). In addition, gene-level synteny analyses were conducted for one-to-one orthologous genes colocalizing to large *A. suum* or *B. malayi* assembly scaffolds (>1 Mb) according to ref. 9. Orthology was determined by pairwise reciprocal BLASTx comparisons between *A. suum*, *B. malayi* and *C. elegans* according to ref. 11. One-to-one orthologous genes shared between either *A. suum* or *B. malayi* and *C. elegans* but not shared among *A. suum* and *B. malayi* based on reciprocal BLASTp analysis were further confirmed by Hidden Markov Modelling using the *jackhmmr* command in the program HMMER 3.0 (ref. 50) and a highly permissive threshold (HMM cutoff: $10^{−2}$).

We assessed the genome-wide variation in the exonic regions by mapping all raw reads from our transcriptomic data to the genomic coding domains using Maq[51], and calling SNPs with a minimum coverage threshold of ten reads. All mapped reads were assessed as synonymous (non-coding change), non-synonymous (coding change) or ambiguous (a SNP that was represented in our data set as an ambiguous IUPAC code wherein one nucleotide change would cause a synonymous mutation and the other a non-synonymous mutation) using a custom Perl script (snp_analysis.pl). All genes were then ranked based on their accumulation of SNPs to assess and identify their levels of conservation/variation relative to their

function. We reasoned that, in addition to the real effects of the variability of each gene on their accumulation of SNPs, these data would be influenced also by the coverage achieved for each gene, which is affected by the number of reads available for each gene (that is, their relative levels of transcription) and the length of each gene. Thus, before ranking, the SNP data for each gene was normalized for its calculated reads per kilobase per million reads (RPKM) and total gene length using the simple equation: SNPs per read per kilobase = total SNPs divided by RPKM divided by gene length (in bp) multiplied by 1,000 bp. Following ranking, we explored function among the 2.5% most variable (with the highest rankings based on normalized SNP data) and most conserved genes (with the lowest rankings based on normalized SNP data). Noting the potential inaccuracy associated with estimating the normalized SNP rankings of lowly transcribed genes (owing to a lack of data/coverage), only genes for which at least 100 reads were available were considered in these functional comparisons.

Functional annotation of coding genes. Following the prediction of the protein-coding gene set, each inferred amino acid sequence was assessed for conserved protein domains in the SProt, Pfam, PRINTS, PROSITE, ProDom and SMART databases using InterProScan[52], employing default settings. Gene ontology categories[53] were assigned to each contig inferred to contain at least one conserved protein domain. Gene ontology categories were summarized and standardized to level 2 and level 3 terms, defined using the GOslim hierarchy[54] using WEGO[55]. To characterize further the contigs/transcripts from *A. suum*, we conducted a series of high-stringency BLASTp homology searches (e-value cut-off: 10^{-5}) against a variety of databases. Each contig was assessed for a known functional orthologue, defined using the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (www.kegg.com). Where appropriate, orthologous matches were mapped visually to a defined pathway using the KEGG pathway tool (available via www.kegg.com) or clustered to a known protein family using the KEGG-BRITE hierarchy tool (available via www.kegg.com). In addition, the amino acid sequence inferred from each *A. suum* coding gene was compared by BLASTp with protein sequences available for key nematode species (*B. malayi*, *C. elegans*, *P. pacificus* and *M. incognita*) as well as for *Drosophila melanogaster*[4] and *Mus musculus*[56] and those contained within the UniProt[57], SwissProt and TREMBL databases[58]. Key protein groups (for example, peptidases, kinases, phosphatases, GTPases, GPCRs, and transport and channel proteins) were characterized by high-stringency BLASTp homology searching (e-value cut-off $<10^{-5}$) of manually curated information sequence data available in the MEROPS[59], WormBase, KS-Sarfari (https://www.ebi.ac.uk/chembl/sarfari/kinasesarfari) and GPCR-Sarfari (https://www.ebi.ac.uk/chembl/sarfari/gpcrsarfari) and the Transporter Classification database[60]. E/S proteins were predicted using Phobius[61], employing both the neural network and hidden Markov models, and by BLASTp homology-searching of the validated signal peptide database[62] and an E/S database containing published proteomic data for *B. malayi*[63,64], *Schistosoma mansoni*[65] and *M. incognita*[66]. In the final annotation, proteins inferred from genes were classified based on a homology match (e-value cut-off: $\leq 10^{-5}$) to: (1) a curated, specialist protein database, followed by (2) the KEGG database, followed by (3) the UniProt/SwissProt/TREMBL databases, followed by (4) the annotated gene set for a model organism, including *C. elegans*, *D. melanogaster*, *M. musculus* or *S. cerevisiae*, followed by (5) the gene ontology classification, and, finally, (6) a recognized, conserved protein domain based on InterProScan analysis. Any inferred proteins lacking a match (BLASTp cut-off $\leq 10^{-5}$) in at least one of these analyses were designated hypothetical proteins. The final annotated protein-coding gene set for *A. suum* is available for download at WormBase (in nucleotide and amino acid formats).

Differential transcription analysis. Following RNA-seq, all paired-end reads for each library constructed were aligned to the predicted *A. suum* gene set using TopHat, and quantitative levels of transcription (RPKM)[67] were calculated using Cufflinks. Differential transcription was assessed[68] using a *P*-value cut-off of ≤0.01 and a minimum, two-fold difference in absolute RPKM values. False discovery rates for differential transcription were determined[68]. To allow the rapid visual assessment of the statistically significant changes in transcription of each gene between and among individual libraries, we constructed heat-maps representing absolute differences in the RPKM values, calculated for each transcript using a customized Perl script (express_heatmap_RPKM.pl). Genetic interaction networks were predicted[69] based on data available for homologous genes in *C. elegans* (inferred from BLASTp comparisons) and viewed using the program BioLayout 3D[70].

Essentiality and druggability predictions. *A. suum* genes with homology to those in the *C. elegans* and/or *D. melanogaster* genomes were inferred based on BLASTp comparisons using the predicted protein sequences for individual species (e-value cut-off 10^{-5}). Phenotypic data for each *C. elegans* and *D. melanogaster* homologue were sourced from WormBase and FlyBase (www.flybase.org), respectively. *A. suum* genes determined[71] to have homologues with lethal phenotypes in both *C. elegans* and *D. melanogaster* were inferred to represent essential

genes. Metabolic chokepoints were defined[29,72] and assessed based on *A. suum* gene sequences determined, by BLASTp comparison (10^{-5}), to have an orthologue in the KEGG database. All 'essential' homologues and/or molecules in 'chokepoints' were then queried against the BRENDA[73] and CHEMBL databases (accessible via https://www.ebi.ac.uk/chembldb/), to identify known chemical inhibitors.

Additional bioinformatic analyses, and use of software. Data analysis was conducted in a Unix environment or Microsoft Excel 2007 using standard commands. Bioinformatic scripts required to facilitate data analysis were designed using Perl, BioPerl, Java and Python and are available via http://research.vet.unimelb.edu.au/gasserlab/.

31. Cantacessi, C. *et al.* Differences in transcription between free-living and CO₂-activated third-stage larvae of *Haemonchus contortus. BMC Genom.* **11,** 266, doi:10.1186/1471-2164-11-266 (2010).
32. Saeed, I., Roepstorff, A., Rasmussen, T., Hog, M. & Jungersen, G. Optimization of the agar-gel method for isolation of migrating *Ascaris suum* larvae from the liver and lungs of pigs. *Acta Vet. Scand.* **42,** 279–286 (2001).
33. Gasser, R. B. *et al.* Single-strand conformation polymorphism (SSCP) for the analysis of genetic variation. *Nature Protocols* **1,** 3121–3128 (2007).
34. Sambrook, J. & Russell, D. W. *Molecular Cloning: A Laboratory Manual* 3rd edn, Vol. 3, E.3–E.4 (Cold Spring Harbor Laboratory, 2001).
35. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2,** 231–239 (1988).
36. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
37. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966–1967 (2009).
38. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
39. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
40. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
41. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protocols Bioinformatics* Ch. 4.10 (2009).
42. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35,** W265–W268 (2007).
43. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21** (Suppl. 1), i152–i158 (2005).
44. 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).
45. She, R., Chu, J. S., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.* **19,** 143–149 (2009).
46. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14,** 988–995 (2004).
47. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
48. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28,** 511–515 (2010).
49. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* **100,** 11484–11489 (2003).
50. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23,** 205–211 (2009).
51. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18,** 1851–1858 (2008).
52. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33,** W116–W120 (2005).
53. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32,** D258–D261 (2004).
54. Camon, E. *et al.* The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.* **13,** 662–672 (2003).
55. Ye, J. *et al.* WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* **34,** W293–W297 (2006).
56. Chinwalla, A. T. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
57. Wu, C. H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34,** D187–D191 (2006).
58. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31,** 365–370 (2003).
59. Rawlings, N. D., Barrett, A. J. & Bateman, A. MEROPS: the peptidase database. *Nucleic Acids Res.* **38,** D227–D233 (2010).
60. Saier, M. H. Jr, Yen, M. R., Noto, K., Tamang, D. G. & Elkan, C. The Transporter Classification Database: recent advances. *Nucleic Acids Res.* **37,** D274–D278 (2009).
61. Kall, L., Krogh, A. & Sonnhammer, E. L. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res.* **35,** W429–W432 (2007).
62. Chen, Y. *et al.* SPD—a web-based secreted protein database. *Nucleic Acids Res.* **33,** D169–D173 (2005).
63. Bennuru, S. *et al.* *Brugia malayi* excreted/secreted proteins at the host/parasite interface: stage- and gender-specific proteomic profiling. *PLoS Negl. Trop. Dis.* **3,** e410 (2009).

64. Hewitson, J. P. *et al.* The secretome of the filarial parasite, *Brugia malayi*: proteomic profile of adult excretory-secretory products. *Mol. Biochem. Parasitol.* **160,** 8–21 (2008).

65. Cass, C. L. *et al.* Proteomic analysis of *Schistosoma mansoni* egg secretions. *Mol. Biochem. Parasitol.* **155,** 84–93 (2007).

66. Bellafiore, S. *et al.* Direct identification of the *Meloidogyne incognita* secretome reveals proteins with host cell reprogramming potential. *PLoS Pathog.* **4,** e1000192 (2008).

67. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).

68. Audic, S. & Claverie, J. M. The significance of digital gene expression profiles. *Genome Res.* **7,** 986–995 (1997).

69. Zhong, W. & Sternberg, P. W. Genome-wide prediction of *C. elegans* genetic interactions. *Science* **311,** 1481–1484 (2006).

70. Goldovsky, L., Cases, I., Enright, A. J. & Ouzounis, C. A. BioLayout(Java): versatile network visualisation of structural and functional relationships. *Appl. Bioinform.* **4,** 71–74 (2005).

71. Doyle, M. A., Gasser, R. B., Woodcroft, B. J., Hall, R. S. & Ralph, S. A. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genom.* **11,** 222, doi:10.1186/1471-2164-11-222 (2010).

72. 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).

73. Scheer, M. *et al.* BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.* **39,** D670–D676 (2011).