3
Integrating and Mining Helminth Genomes to Discover and Prioritize Novel Therapeutic Targets

Dhanasekaran Shanmugam, Stuart A. Ralph, Santiago J. Carmona, Gregory J. Crowther, David S. Roos, and Fernán Agüero[*]

Abstract

Diseases caused by helminth parasites remain the most neglected of tropical diseases. Consequently, discovery of new therapeutics, diagnostics, and vaccines for helminth parasites has been slow to progress. This is in part because anthelmintic discovery has relied upon biological screens – either genetic or chemical – in model as well as in parasitic worms and these approaches are limited in a number of ways. For example, genetic manipulation (such as RNA interference) is still not available for many helminth parasites, and, therefore, genomic-scale experimental target identification and validation studies remain challenging. Also, for many parasitic helminths, the life cycle of the parasite cannot be maintained *in vitro*, thus limiting experimental screens. To facilitate discovery of new targets, a genomics-based approach has been gaining traction and will be supported by the increasing numbers of complete genome sequences available for helminth parasites. The availability of these genome sequences is expected to support a wide variety of genomic-scale studies that will generate functional datasets regarding the expression, structure, phylogeny, essentiality, and validation of genes from parasite stages relevant to disease. In addition, target identification will be facilitated by mapping of functional data through orthology from model organisms like *Caenorhabditis elegans*. In order to realize the full potential of genomics-based target discovery, various functional datasets need to be integrated with genome sequence information in a structured format that can be easily accessed and mined for anthelmintic target discovery. This chapter discusses advances in genomics-driven target discovery for helminths, and highlights the increasing contribution of data repositories such as TDR Targets and WormBase to anthelmintic target discovery. The search strategies implemented in the TDR Targets database will be used to illustrate the utility of comparative genomics to discover potential helminth drug targets and identify missing functional datasets in helminth parasites that will greatly improve target identification.

[*] Corresponding Author

Parasitic Helminths: Targets, Screens, Drugs and Vaccines, First Edition. Edited by Conor R. Caffrey

Introduction

In the era of genome sequencing, it is inevitable that comparative genomics plays a major role in driving biological and therapeutic discoveries. This is especially true for less-studied organisms, like the helminth parasites, that are not easily amenable to experimental manipulation in the laboratory. Helminth parasites (nematodes, cestodes, and trematodes) are a diverse group of invertebrate animals with complex life cycles during which they infect various animal hosts. The disease burden resulting from human helminth infections is enormous. Estimates from the World Health Organization (WHO) indicate that more than 2 billion people world-wide are at risk of acquiring such infections (http://www.who.int/tdr/svc/diseases/helminths). Few effective drugs are available to treat and control helminth infections [1, 2] (see also Chapters 14 and 20), and there are already a significant number of documented cases of resistance in veterinary helminths [3, 4]. In addition, recent reports suggest that resistance to some anthelmintic treatments in humans might be emerging [5, 6]; thus, alternate therapeutics are urgently required.

Although whole-organism (phenotypic) screening approaches have dominated the drug discovery landscape for helminth parasites, target-based approaches are gaining ground, being supported by a number of advances (see also Chapters 1 and 8). (i) There is the increased availability of genome sequence information for a number of helminth parasites, which is supported by parasitic helminth genome initiatives such as those of the Wellcome Trust Sanger Institute and Washington University's Genome Sequencing Center [7, 8]. (ii) There has been a steady increase in genome-wide studies on expression profiling (primarily by expressed sequence tag (EST) library sequencing and microarray analysis [9–12]), proteomics [13–15], and validation of function and phenotype (most often by RNA interference (RNAi) [16]; see also Chapters 6 and 7). In fact, recent advances in deep sequencing of both genomic and mRNA mean that there will not be a shortage of sequence data for helminth parasites. Rather, the focus is now shifting towards developing methods and tools to effectively integrate and use these datasets in order to understand the biology of helminths and support anthelmintic discovery.

Traditional genome repositories such as organism-specific genome databases and the National Center for Biotechnology Information GenBank [17] serve the essential function of hosting raw genome sequence data and associated genomic or gene-specific annotations, generated mostly by standardized computational pipelines. With the increasing availability of different types of genomic-scale experimental datasets, it is now standard practice by most genome servers to integrate these datasets and allow end-users to query the available data in order to retrieve desired genes. Moreover, genome servers such as GeneDB [18] and EuPathDB [19], which host genome data for various eukaryotic pathogens, have implemented tools to perform comparative genomics analysis of related species. Comparative genomics is especially important in the case of helminth pathogens for which the functional annotation of available genomes suffers from a relative lack of robust experimental tools. For example, the highly annotated genome sequence information available for *Caenorhabditis elegans* in WormBase [20] can be used to inform gene annotation for orthologs in parasitic helminths. In fact, this strategy has been already used to identify potential target genes in *Brugia malayi* [21] and *Schistosoma mansoni* [22]. Implementing a similar workflow, but as part of a curated database, will allow for the continual updating of underlying datasets.

We outline how *in silico* comparative genomics can be employed to enhance our understanding of helminth biology and assist in the discovery of novel drug and vaccine targets. Specifically, grouping proteins by orthology [23] has been useful to map functional genomic datasets such as metabolic pathways and genetic phenotypes from the model organism, *C. elegans*, to parasitic helminths. This chapter also reviews how genome-wide annotation that is integrated into genome databases can be used to identify and prioritize target genes. Finally, we provide a brief overview of the WHO's TDR Targets database [24], which integrates a number of datasets mapped to the genomes of different pathogens, including helminths, and provides the necessary informatics tools to prioritize target genes. Illustrative examples of target prioritization for *B. malayi*, *S. mansoni*, *Onchocerca volvulus*, and *Wuchereria bancrofti* will be demonstrated using the tools implemented in TDR Targets.

Availability of Genome Sequence Information for Parasitic Helminths

In addition to the model nematode, *C. elegans* [25], which was the first nematode genome to be sequenced, a handful of other parasitic worm genomes, including *B. malayi* [26], *S. mansoni* [27], and *Trichenella spiralis* [28], have been sequenced. Table 3.1 provides a list of worms, most of them parasitic in animals, which are currently under study by various sequencing centers. However, there are significant challenges ahead in terms of producing high quality genome annotation, making data accessible to the community, and enabling functional studies. Genome repositories and databases will be important here. For helminths, the Nematode.Net database [29] maintains a collection of sequences, both genomic and EST-based, and provides various functionalities such as functional classification, ortholog identification, and expression data analysis. Although this database also includes data for *C. elegans*, more sophisticated phenotype data (based on targeted gene disruption or RNAi studies) for *C. elegans* can be mined from WormBase [20] and used to identify potential targets in parasitic helminths (see Figure 3.1). Also, the GeneDB [18] and SchistoDB [30] databases provide access to genome sequence annotation for a number of *Schistosoma* species. The TDR Targets database contains genome information for *B. malayi* and *S. mansoni*, and integrates a variety of datasets (see below), including orthology-based mapping of *C. elegans* phenotype data, to aid in the identification of potential drug targets. The TDR Targets database will incorporate genome information for other parasitic worms as data become available.

Table 3.1 Sequence data availability for helminth organisms (as of March 2012).

Organism type	Number of species with EST data[a]	Number of species with genome sequence data[a]	Number of species with RNA sequence data[a]
Trematoda – flukes	3	2 (Sman; Sjap)	2 (Sman; Sjap)
Cestoda – tapeworm	2	4 (Egra; Emul; Hmic; Tsol)	3 (Egra; Emul; Hmic)
Nematoda – clade I	4	2 (Tspi; Tmur)	—
Nematoda – clade III	6	4 (Bmal; Asuu; Alum; Ovol)	1 (Asuu)
Nematoda – clade IV a/b	18	4 (Srat; Gpal; Minc; Hgly)	2 (Srat; Gpal)
Nematoda – clade V	13	13 (*Caenorhabditis* spp;[b] Acan; Acey; Aduo; Conco; Dviv; Name; Oden; Oost; Hcon; Nbra; Tcir; Ppac)	3 (*Caenorhabditis* spp;[b] Anca: Conc; Dviv; Hbac; Name; Nbra; Oden; Oost; Tcir; Tcol)

The table summarizes the various helminth species (both parasitic and nonparasitic) for which EST, genome, and RNA sequence data is either available or will soon be available. Sman, *Schistosoma mansoni*; Sjap, *Schistosoma japonicum*; Egra, *Echinococcus granulosus*; Emul, *Echinococcus multilocularis*; Hmic, *Hymenolepis microstoma*; Tsol, Taenia solium; Tspi, *Trichinella spiralis*; Tmur, *Trichuris muris*; Bmal, *B. malayi*; Asuu, *Ascaris suum*; Alum, *Ascaris lumbricoides*; Ovol, *Onchocerca volvulus*; Srat, *Strongyloides ratti*; Gpal, *Globodera pallida*; Minc, *Meloidogyne incognita*; Hgly, *Heterodera glycines*; Acan, Ancylostoma caninum; Acey, Ancylostoma ceylanicum; Aduo, Ancylostoma duodenale; Conco, Cooperia oncophora; Dviv, Dictyocaulus viviparus; Hbac, Heterorhabditis bacteriophora; Name, Necator americanus; Oden, Oesophagostomum dentatum; Oost, Ostertagia ostertagi; Tcir, Teladorsagia circumcincta; Tcol, Trichostrongylus colubriformis; Hcon, *Haemonchus contortus*; Nbra, *Nippostrongylus brasiliensis*; Tcir, *Teladorsagia circumcincta*; Ppac, *Pristionchus pacificus*.

a) Indicates both completed and projects in progress; data obtained from Nematode.Net, the Wellcome Trust Sanger Institute, the NCBI Genome and Gene Expression Omnibus Databases, and [8].

b) Includes multiple species of *Caenorhabditis*.

Overview of Genome Annotation Datasets that Aid Target Identification

In addition to storing the sequence information for any given genome, the respective genome databases also provide a collection of annotations describing several different features either on a genomic-scale or in a gene/protein-specific manner. Some of the annotations are gathered automatically (e.g., identification of open reading frames and their protein domains/properties to predict function), but many others are culled from experimental data and have to be constantly curated as the data become available. The latter set includes data that describe gene/protein expression and regulation, genetic variations, protein structure, gene essentiality, phenotypic/functional responses to genetic/chemical alterations of gene structure or function, ligand/inhibitor interactions, and any other relevant information (see Table 3.2 for a list of datasets that can be mapped to a sequenced genome). Although the data available from automated genome annotations are very similar in format and accessibility across various genomes, the quality and depth of coverage of experimental data varies widely between organisms. This is especially true for helminths in that the data available for *C. elegans* are much more comprehensive than those for parasitic worms. This is, in part, due to the patchy genome information available for parasitic worms, but also due to them

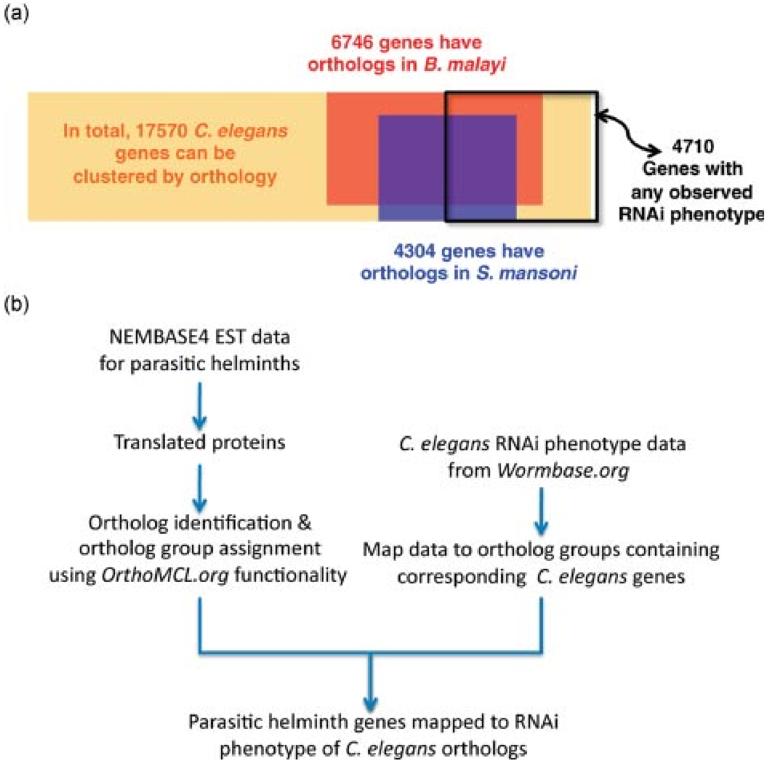
(a)
6746 genes have orthologs in B. malayi
In total, 17570 C. elegans genes can be clustered by orthology
4710 Genes with any observed RNAi phenotype
4304 genes have orthologs in S. mansoni
(b)
NEMBASE4 EST data for parasitic helminths
Translated proteins
Ortholog identification & ortholog group assignment using OrthoMCL.org functionality
C. elegans RNAi phenotype data from Wormbase.org
Map data to ortholog groups containing corresponding C. elegans genes
Parasitic helminth genes mapped to RNAi phenotype of C. elegans orthologs

Figure 3.1 Mapping RNAi phenotype data using orthology. (a) The illustration shows how a significant number of *C. elegans* genes with orthologs in *B. malayi* and *S. mansoni* are also associated with at least one observed RNAi phenotype. Using this mapping information, desirable phenotypic data can be used to select for target genes from parasites. (b) General informatics pipeline used for transient mapping of phenotypic data from *C. elegans* to parasitic organisms.

being less experimentally tractable than *C. elegans*. Thus, it is useful to implement orthology-based transient mapping of annotation across closely related species to identify putative target genes, as described in the following section.

Orthology-Based Annotations and Comparative Genomics

Orthologs are defined as homologous proteins that are separated by a speciation event and are considered to be functionally conserved across species. Orthologous proteins may be estimated by reciprocal (two-way) best hits using BLAST. Precomputed ortholog pairs from more than 100 different species have been clustered into ortholog groups which can be accessed from the OrthoMCL database [23]. For example, Figure 3.1a illustrates how the RNAi phenotype data available for *C. elegans* genes can be mapped to orthologs in *B. malayi* and *S. mansoni*. Of the 4710 genes with observed RNAi phenotypes in *C. elegans* (available in WormBase [20]), 2242 are

Table 3.2 Genome annotation data made available through databases.

Annotation type	Annotated data
Automatic annotation and sequence based predictions	Gene ID
	Product name
	Gene/protein sequence
	Protein length in amino acids
	Molecular weight
	Isoelectric point
	Transmembrane domain
	Signal/transit sequences
	Protein domain by blast similarity
	Gene ontology predictions
	Pathway mapping and enzymes
	Metabolite mapping
	Protein structure model
	Phylogeny and orthology
	Druggability
Experimental evidence based annotation	Expression: anatomical and life cycle stage specificity
	Enzymatic activity and kinetics
	Metabolite and inhibitor ligand binding
	Recombinant availability
	Protein structure data
	Gene essentiality (knockout/down): life cycle stage specificity
	Phenotype (Genetic/Chemical): life cycle stage specificity

A variety of annotations, either from sequence-based prediction or based on experimental evidence, are made available through genome databases. For parasitic helminths in particular, the lack of genome wide experimental datasets needs to be addressed.

orthologous to 3064 *Schistosoma* genes and 3058 are orthologous to 3377 *Brugia* genes, with a large degree of overlap between orthologs of the two parasites. Similarly, the other annotations listed in Table 3.2 can be transiently mapped onto the genome of interest using information from a suitable model organism. However, the approach has some drawbacks. For example, whereas mapping enzymes and metabolic pathways using orthology is most likely to be correct, mapping genetic essentiality data can be misleading as a large proportion of these data tends to be organism-specific (see [31] and references therein). Therefore, in using such datasets, one needs to be familiar with the biology of each species and the suitability of the mapped data for a particular organism. Nevertheless, comparing orthologous genes across species is important when performing comparative genomics.

In addition to transient annotation of genomic-scale data as discussed above, the identification of orthologs enables phylogenetic profiling of the genome of interest. The presence or absence of genes in the host versus parasite genomes often provides a first-stage filter to narrow down the set of target genes of interest for further analysis [32]. As an example, one may want to select *B. malayi* genes that are absent from free-living nematodes, but are present in parasitic nematodes. Such a selection is likely to enrich for genes that are essential for parasitism and hence of interest as targets. Ortholog clustering also helps to identify gene duplications that can contribute to functional redundancy and genetic variation, sometimes even within different strains (isolates) of the same species [33]. Therefore, implementing orthology-based querying as part of database infrastructure can be of tremendous use to select target genes with the desired properties. TDR Targets implements this functionality making use of ortholog groupings already available in the OrthoMCL database. In the following sections, examples of target identification through orthology for various parasitic helminths are presented.

Predictions of Essentiality

A major task in the genome-wide prediction of promising drug targets for anthelmintics is the identification of essential genes [34, 35]. Essential targets are those for which inhibition of protein function is most likely to result in death, a severe phenotype(s), or a significant loss of fitness. These are not the only imaginable targets of anthelmintics. Indeed, chemical modulation of nonessential genes is a well-trodden path to eliminating helminths, such as agonists of nonessential genes that induce loss of muscle control and, therefore, worm expulsion (see Chapter 14 for examples). Identification of essential genes does enjoy one advantage in that resistance will not arise due to loss of function or deletion – resistance to some drugs in other pathogens involves deletion of nonessential genes including melarsaprol resistance in *Trypanosoma brucei* [36] and capreomycin resistance in *Mycobacterium tuberculosis* [37].

Although advances are being made in reverse genetics tools for pathogenic helminths, working either on a genome-wide scale or at the level of the individual gene, these are either lacking or rudimentary. Therefore, inference of essentiality through bioinformatic means is desirable and several potential methods are available, as discussed below. *C. elegans* has been employed as a model to suggest essentiality in important pathogenic nematodes, such as *B. malayi* [21], and it may also be useful to determine essentiality in *Onchocerca* spp. [38] and *Strongyloides* spp. [39]. Also, and as discussed above, TDR Targets allows the identification of *Brugia* genes with essential orthologs in *C. elegans*. The caveats to such inferences are that the parasitic lifestyle may allow these species to dispense with genes that are essential in free-living nematodes, whereas other molecules involved in host–pathogen interactions become newly essential. For helminths other than nematodes, it is less obvious whether essentiality can be inferred from functional data for *C. elegans* alone, and whether other animal models, such as the fruitfly, *Drosophila melanogaster*, can (should) be incorporated. A case in point is the analysis by Caffrey *et al.* [22] that filtered gene disruption data for *both C. elegans* and *D. melanogaster* to predict essential *S. mansoni* genes.

Helminth genes most likely to be essential are those that are shared by a greater number of evolutionarily diverse organisms [40]. One strategy, therefore, to determine essentiality in parasitic helminths is to focus on just these "essential" helminth genes that have orthologs in other phyla and are supported, where possible, by experimental data actually demonstrating essentiality. An unfortunate corollary of this is that helminth genes that are absent from the human host are less likely to be essential than those that are shared. Maximizing prediction of essentiality by choosing evolutionarily conserved proteins may therefore conflict with eventually developing ligands (inhibitors) that selectively target parasite proteins. In practice, therefore, these conflicting criteria must be balanced to identify targets that can be selectively drugged, but which are still likely to be essential. Also, as discussed by Caffrey *et al.* [22], it is often the case that selectivity and potency of any ligand eventually comes down to a range of parasitological and physiological factors, and smart medicinal chemistry to avoid "off-target" toxicity to the host.

An alternative to simply porting essentiality from experimentally characterized orthologs onto helminth genes is to rank essentiality based on gene product properties that are predictive of essentiality. One such approach is essentiality prediction from network connectivity. Proteins with larger numbers of interactions – called hubs – are more likely to be essential across several eukaryote groups [41, 42]. This observation informed the successful prediction and verification of essentiality in nematodes using highly connected genes in WormNet, which is a network that incorporates protein–protein interactions as well as other data types including coexpression, co-occurrence of gene names in text, and genetic interactions [43]. Future systems biology data from parasitic helminths could potentially be integrated into similar networks to improve and extend such network-based predictions.

Another property of gene products that can be used to infer essentiality is their position in metabolic networks. A number of genome-wide methods such as chokepoint analyses are available to predict essentiality and the curated *Schistosoma* metabolic network Schistocyc [30] or the NemaPath mapping of KEGG pathways [44] are useful starting templates for such analyses that will hopefully be replicated in other parasitic helminths.

Orthology-Based RNAi Phenotype Data Mapping Between *C. elegans* and Parasitic Helminths

C. elegans has been an important model organism to inform both experimental and comparative genomics studies with various parasitic helminths (see also Chapter 2). Examples of target identification for both *B. malayi* and *S. mansoni* using phenotype data available for *C. elegans* are published [20, 22, 45]. Here, we illustrate for *Onchocerca* spp. and *Wuchereria bancrofti* how this approach can be applied even in the absence of complete genome information (Figure 3.1b). In order to perform this analysis, EST data for various *Onchocerca* spp. and *W. bancrofti* were obtained from the NEMBASE4 database [46] that hosts EST data from more than 60 nematode species. The translated protein sequences for these ESTs were then used to identify the corresponding *C. elegans* orthologs using the ortholog identification pipeline implemented at the OrthoMCL database [23]. RNAi phenotypes for *C. elegans* genes obtained from the WormBase database were then mapped to the relevant ortholog groups of *Onchocerca* spp. and *W. bancrofti*. A supplementary file containing the results from this mapping exercise is available from TDR Targets using the link http://tdrtargets.org/static/shanmugam-helminth-genomes/Table-III.xlsx. This file lists all ortholog groups that contain at least one gene in *C. elegans* that has an observed RNAi phenotype and at least one gene in *Onchocerca* or *W. bancrofti*. From this list, parasite genes mapped to select RNAi phenotypes can be identified and pursued further as potential targets. As ever, downstream experimental work is required to validate the essentiality of these targets in these species. Once the genome sequence of these parasites becomes integrated into the TDR Targets database, genomic-scale mapping of the above exercise can be carried out. Using the workflow and functionalities implemented in TDR Targets (see Figure 3.2 and the discussion

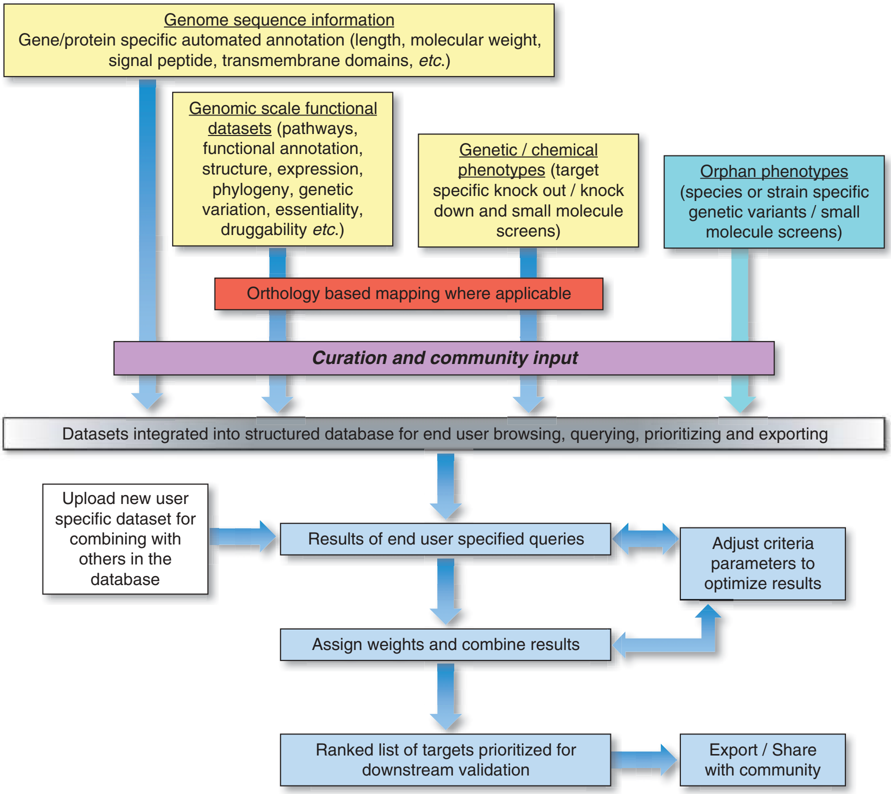
Genome sequence information
Gene/protein specific automated annotation (length, molecular weight, signal peptide, transmembrane domains, *etc.*)
Genomic scale functional datasets (pathways, functional annotation, structure, expression, phylogeny, genetic variation, essentiality, druggability *etc.*)
Genetic / chemical phenotypes (target specific knock out / knock down and small molecule screens)
Orphan phenotypes (species or strain specific genetic variants / small molecule screens)
Orthology based mapping where applicable
Curation and community input
Datasets integrated into structured database for end user browsing, querying, prioritizing and exporting
Upload new user specific dataset for combining with others in the database
Results of end user specified queries
Adjust criteria parameters to optimize results
Assign weights and combine results
Ranked list of targets prioritized for downstream validation
Export / Share with community

Figure 3.2 TDR Targets database structure and workflow for prioritizing genes. This scheme illustrates how various datasets integrated into the TDR Targets database can be used to identify and prioritize target genes. Note that mapping data from orthologous genes in other species, curation of published data, and community input on selected targets are key to making this process work. TDR Targets has implemented a user-friendly database infrastructure and easy-to-use informatics tools, all of which aid in target identification. For details of database functionality and case scenarios, see [24, 45].

below), we will next discuss the utility of genomic-scale datasets to identify targets of interest.

The TDR Targets Database

TDR Targets facilitates target identification for major tropical pathogens, including *B. malayi* and *S. mansoni* [24]. It contains genome information for all the pathogens listed in the target search page, and also integrates a variety of functional datasets (see Table 3.2) that facilitate the formulation of user-defined queries [45]. The various informatics tools provided via an open-access web interface, allow users to browse and query data, view and modify results, rank genes based on user-assigned weight values for selected criteria, export data and share results. TDR Targets obtains genome information for each species from various genome repositories like EuPathDB, GeneDB, WormBase, and GenBank. Functional annotations are obtained by a combination of methods, including orthology-based mapping of data across species, curation of literature information, and generating in-house datasets in collaboration with academic and industry partners. Figure 3.2 shows the general workflow of how one may carry out a target selection exercise using TDR Targets. First, the user searches for targets in the selected pathogen by formulating one or more queries based on the available datasets. The result of these queries can then be viewed as a list and exported as text or Excel files. The user can also manage queries from the *history* page by combining them using the *union* and *intersection* functions or *delete, export, rename,* and view the criteria used to formulate the queries. As an option, registered users can save the queries and publish the prioritized list of genes on the website to share with others.

Two different strategies can be employed to prioritize target genes using TDR Targets (Figure 3.3). The first is to use the *intersection* functionality to combine the results from multiple queries. As shown in Figure 3.3, five different hypothetical queries are displayed for a helminth genome. Combining in this manner is progressively restrictive because with the intersection of each query many genes are filtered away. Thus, starting from the whole genome of a helminth, which may contain more than 2×10^4 genes, one may end up with less than 100 genes. All genes contained in the resulting list will qualify for all of the criteria employed. Although this method helps to generate a list containing only the desired targets, it has the drawback of excluding genes that failed only one of the criteria used and does not provide the flexibility of modifying the target list without dramatically changing the query parameters or the query itself. One of the main issues affecting intersection-based strategies is the quality of the available genomes and their annotation. If resources allocated to manual curation of a genome are limited, or if the body of experimental evidence for any given genome is not sufficiently large or diverse, then it is more likely that many genes will fail to meet simple criteria that depend on the quality of annotation (e.g., “kinase” will not match a kinase that was annotated as “hypothetical protein”). Poor gene identification strategies (incorrect gene models that lead to the wrong identification of translational start sites and/or splicing sites)

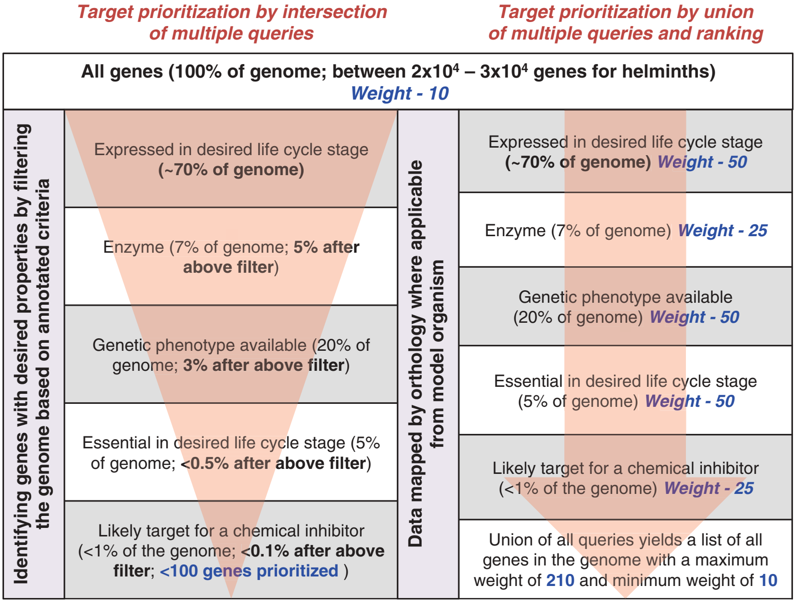
Target prioritization by intersection of multiple queries
Target prioritization by union of multiple queries and ranking
All genes (100% of genome; between 2x10⁴ – 3x10⁴ genes for helminths)
Weight - 10
Identifying genes with desired properties by filtering the genome based on annotated criteria
Expressed in desired life cycle stage (~70% of genome)
Enzyme (7% of genome; 5% after above filter)
Genetic phenotype available (20% of genome; 3% after above filter)
Essential in desired life cycle stage (5% of genome; <0.5% after above filter)
Likely target for a chemical inhibitor (<1% of the genome; <0.1% after above filter; <100 genes prioritized)
Data mapped by orthology where applicable from model organism
Expressed in desired life cycle stage (~70% of genome) Weight - 50
Enzyme (7% of genome) Weight - 25
Genetic phenotype available (20% of genome) Weight - 50
Essential in desired life cycle stage (5% of genome) Weight - 50
Likely target for a chemical inhibitor (<1% of the genome) Weight - 25
Union of all queries yields a list of all genes in the genome with a maximum weight of 210 and minimum weight of 10

Figure 3.3 Examples of different strategies employed to prioritize targets using the TDR Targets database. The left side demonstrates the use of a more restrictive *intersection* (AND) functionality to run queries, whereas the right side demonstrates the *union* (OR) functionality. When using the *intersection* query, only genes that have qualified for all the selected criteria are obtained. In contrast, the *union* query facilitates a genome-wide ranking that is based on assigned weights for search criteria. For more details, see [24].

can also lead to failures of many downstream bioinformatics predictions (e.g., orthology detection, and domain and motif identification). For many helminth genomes available as drafts, alternative prioritization strategies may help to lower the impact of some of these knowledge gaps.

The second way to prioritize genes is to apply the *union* functionality in combination with weight assignments for individual queries. In Figure 3.3, this is demonstrated using the same queries used above with the intersection example. Note that for each query a *weight* value is assigned and when a gene qualifies for two or more query criteria, the individual weight values add up and provide a way to rank genes. Thus, in the example shown, genes that qualify for all of the selected criteria used to run the queries will receive the maximum weight value of 210 while other genes will receive lower weight values based on their qualifying criteria. The resulting list will contain all the genes from the genome ranked according to their weight values. The ranked target list can be easily modified by adjusting the weight values assigned to each criterion. The ranked list also provides users with a genome-wide perspective on how useful the chosen criteria are for the purpose of target selection. By default, TDR Targets performs a union of multiple queries run by users and provides a ranked list. Alternatively, users can manage and combine their queries in various ways using the functionalities available on the *history* page.

Target Prioritization in *B. malayi* and *S. mansoni* Using the TDR Targets Database

Genomic data for *B. malayi* and *S. mansoni* is integrated into the TDR Targets database and, based on orthology, their genes have been mapped to *C. elegans* phenotypic data. Using these data, examples of target prioritization were carried out for both these parasites [45] and the results are available for viewing and modification on the database site (*B. malayi*, http://tdrtargets.org/published/browse/361; *S. mansoni*, http://tdrtargets.org/published/browse/336). In these examples, targets have been prioritized based on a number of features; phenotype upon RNAi of the *C. elegans* ortholog, availability of structural models, availability of orthologs in *C. elegans*, predicted druggability, function as a catalyst (i.e., an enzyme), and assayability. The weights used for each criterion are heavily biased towards loss-of-fitness phenotypes with less weight for other features. Owing to the differences in the availability of data, but also because we wanted to illustrate the flexibility of the TDR Targets resource, we used somewhat different sets of criteria for each species. One of the main differences was the availability of gene expression data for *S. mansoni*. These data were derived from stage-specific EST sequencing projects available at SchistoDB [30] and were used to give an additional score to those gene products expressed in those life cycle stages relevant to infection in humans.

As a number of current anthelmintics modulate neuromuscular function [47], another useful prioritization strategy may take into account not just the timing of expression (developmentally regulated genes), but also the anatomical location of expression (spatially regulated genes). Genome-wide experimental datasets containing this information are currently lacking for parasitic helminths. Therefore, we used *C. elegans* expression data mapped to the corresponding orthologs of *B. malayi* and *S. mansoni*. These data, derived from a large compendium of microarray analyses (916 experiments from 53 datasets), were recently reanalyzed [48] to obtain subsets of genes that are differentially expressed in various tissues. The underlying hypothesis is that a gene that is expressed in a defined tissue (e.g., muscle) in one organism is more likely to have the same pattern of expression in another related organism. The evolutionary divergence of *B. malayi* and, especially, *S. mansoni* from *C. elegans* will need to be considered accordingly when weighting these criteria.

For this exercise, we used the same criteria and weights as before for *B. malayi* and *S. mansoni* [45], but added additional weights to those genes for which orthologs in *C. elegans* are expressed in nervous or muscular tissues. The results of these prioritizations are available from TDR Targets (*S. mansoni*, http://tdrtargets.org/published/browse/394; *B. malayi*, http://tdrtargets.org/published/browse/395). Although both the previous prioritization exercises and these latest revisions are very similar in displaying cytoskeleton and motor proteins, such as β-tubulin (the target of benzimidazole anthelmintics), dyneins, and myosins, at the top of these lists, a number of additional interesting targets emerge based on the criterion of tissue expression. For *S. mansoni*, among the 37 genes that were raised into the top 100 are a number of potentially druggable targets such as a putative Ras-like GTPase (Smp_146600), a putative calcium-dependent protein kinase (Smp_011660.2), and a putative tyrosine kinase (Smp_136300). For *B. malayi*, a similar approach led to higher scores for a number of potentially druggable targets, including a putative adenylate kinase (Bm1_24575), a Ser/Thr protein phosphatase family (Bm1_41290), and a short-chain dehydrogenase potentially involved in the metabolism of steroids (Bm_45995). Although these targets await further experimental validation, the underlying idea behind these computational exercises is that increased usage of experimental data providing information on different independent criteria (i.e., orthogonal, see [35]) should drive these prioritizations.

Currently Unavailable Genomic Datasets that will Improve Target Prioritization for Parasitic Helminths

A number of key datasets that would enhance the prioritization of targets in parasitic helminths are not yet available. Some of these datasets are high on the lists of priorities of many scientists and funding agencies, and, accordingly, deserve to be listed again. There is a notable lack of genomic-scale assessment of phenotypes caused by either targeted gene disruption or RNAi, particularly for flatworms, since *C. elegans* is not a perfect model of their biology. Targeted gene disruptions (e.g., knockouts) are a more reliable indicator of phenotype than RNAi, for which off-targeting is a real problem [22, 49]. Datasets that reveal the temporal and spatial expression of genes would also be highly valuable. Although substantial transcriptomic sequence information has become available for schistosomes over the last decade (cited in [16]; see also SchistoDB), and high-throughput sequencing has been (and will be) a boon for trematodes and nematode parasites (see also Chapters 4 and 5), the breadth and depth of these datasets requires further improvement. In addition, it is worth noting that protein structure information for these organisms is also limited. Although this is not strictly a validating criterion, knowledge of the structure of a target helps in a number of downstream analyses, such as the identification of potential ligand binding sites, the assessment of the likelihood of binding by small molecules and in the rational design of inhibitors. As of July 2011, the Protein Data Bank [50] carries the following structural data for helminths: Nematoda, 194 structures; Platyhelminths, 71 structures (Cestoda, two structures; Trematoda, 64 structures). Even when taking into account potential redundancies (several structures solved for the same protein), these figures are less than those for other parasites that cause neglected diseases (e.g., 552 and 564 structures available for trypanosomatids and apicomplexans, respectively). Finally, a recent addition to the TDR Targets database is the integration of chemical datasets (available as of Version 4) [51]. The availability of links between targets and compounds, manually curated from the literature, as well as the ability to perform similarity searches between compounds, opens the door to more comprehensive prioritizations. In this context, datasets from high-throughput chemical and whole-organism screens would be valuable as they would allow users to identify chemical scaffolds that are bioactive against helminths and that may then be linked by similarity to other compounds, and ultimately, to potential targets. Furthermore, the inclusion and integration of data for inactive compounds would be as important, not least in avoiding unnecessary duplication of effort. A comparison of the activity profile of any given compound against different parasitic helminths can help to shed light on the potential mode(s) of action of the compound.

Conclusions

Helminth infections are among the most neglected of human diseases relative to their global burden. Despite limited resources dedicated to either fundamental research or applied drug discovery for these organisms, recent whole-genome projects and transcriptomic surveys of many helminths offer promising starting points for chemotherapeutic or vaccine-based interventions. A challenge for these genome projects is that some of the respective organism-specific research communities are relatively small. This means that the human resources and biological technologies to fully exploit the data arising from some helminth genomes are quite limited. This has two important implications. (i) Limited resources make it all the more important to prioritize the most promising therapeutic targets from the myriad of potential macromolecules to work on. (ii) Where possible, relevant chemical and genetic data should be identified and ported from more thoroughly studied organisms. These tasks are limited by the appropriateness of the model organism and organism-specific features that can render cross-species inferences unsound. Nevertheless, genomics-based approaches will facilitate the process of identifying tractable drug targets and finding promising chemical leads for anthelmintic development. Informatic tools such as TDR Targets serve a useful function in organizing the genomic, phyletic, phenotypic, and chemical resources necessary for target identification. Also, cross-organism platforms that combine various individual genome projects (such as EuPathDB, GeneDB, and TDR Targets), assisted by comparative genomics, will certainly provide valuable insights into the biology of helminths and other parasites, particularly those outside the field of tropical diseases. Much remains to be done to alert such potential users to the significance of genomics in helminth drug development and to kindle their interest therein.

Acknowledgments

Development of the TDR Targets database is funded by the Special Programme for Research and Training in Tropical Diseases (WHO/World Bank/UNDP/UNICEF). S.A.R. is an Australian Research Council Future Fellow (FT0990350). S.J.C. and F.A. are fellows of the National Research Council (CONICET), Argentina.

References

1 Holden-Dye, L. and Walker, R.J. (2007) Anthelmintic drugs, in *WormBook* (ed. The *C. elegans* Research Community), doi: 10.1895/wormbook.1.143.1.

2 Nwaka, S. and Hudson, A. (2006) Innovative lead discovery strategies for tropical diseases. *Nat. Rev. Drug Discov.*, **5**, 941–955.

3 Prichard, R.K. (1990) Anthelmintic resistance in nematodes: extent, recent understanding and future directions for control and research. *Int. J. Parasitol.*, **20**, 515–523.

4 Kaplan, R.M. (2004) Drug resistance in nematodes of veterinary importance: a status report. *Trends Parasitol.*, **20**, 477–481.

5 Osei-Atweneboana, M.Y., Eng, J.K.L., Boakye, D.A., Gyapong, J.O., and Prichard, R.K. (2007) Prevalence and intensity of *Onchocerca volvulus* infection and efficacy of ivermectin in endemic communities in Ghana: a two-phase epidemiological study. *Lancet.*, **369**, 2021–2029.

6 Osei-Atweneboana, M.Y., Awadzi, K., Attah, S.K., Boakye, D.A., Gyapong, J.O., and Prichard, R.K. (2011) Phenotypic evidence of emerging ivermectin resistance in *Onchocerca volvulus. PLoS Negl. Trop. Dis.*, **5**, e998.

7 Berriman, M., Lustigman, S., and Mc Carter, J.P. (2007) Helminth initiative for drug discovery – report of the informal consultation, genomics and emerging drug discovery technologies. *Expert Opin. Drug Discov.*, **2**, S83–S89.

8 Brindley, P.J., Mitreva, M., Ghedin, E., and Lustigman, S. (2009) Helminth genomics: the implications for human health. *PLoS Negl. Trop. Dis.*, **3**, e538.

9 Parkinson, J., Anthony, A., Wasmuth, J., Schmid, R., Hedley, A., and Blaxter, M. (2004) Partigene – constructing partial genomes. *Bioinformatics*, **20**, 1398–1404.

10 Parkinson, J., Mitreva, M., Whitton, C., Thomson, M., Daub, J., Martin, J., Schmid, R., Hall, N. *et al.* (2004) A transcriptomic analysis of the phylum Nematoda. *Nat. Genet.*, **36**, 1259–1267.

11 Ramanathan, R., Varma, S., Ribeiro, J.M.C., Myers, T.G., Nolan, T.J., Abraham, D., Lok, J.B., and Nutman, T.B. (2011) Microarray-based analysis of differential gene expression between infective and noninfective larvae of *Strongyloides stercoralis. PLoS Negl. Trop. Dis.*, **5**, e1039.

12 Wasmuth, J., Schmid, R., Hedley, A., and Blaxter, M. (2008) On the extent and origins of genic novelty in the phylum Nematoda. *PLoS Negl. Trop. Dis.*, **2**, e258.

13 Marcilla, A., Sotillo, J., Pérez-Garcia, A., Igual-Adell, R., Valero, M.L., Sánchez-Pino, M.M., Bernal, D., Muñoz-Antolí, C. *et al.* (2010) Proteomic analysis of *Strongyloides stercoralis* l3 larvae. *Parasitology*, **137**, 1577–1583.

14 Mulvenna, J., Hamilton, B., Nagaraj, S.H., Smyth, D., Loukas, A., and Gorman, J.J. (2009) Proteomics analysis of the excretory/secretory component of the blood-feeding stage of the hookworm, *Ancylostoma caninum. Mol. Cell Proteomics*, **8**, 109–121.

15 Weinkopff, T., Atwood, J.A., Punkosdy, G.A., Moss, D., Weatherly, D.B., Orlando, R., and Lammie, P. (2009) Identification of antigenic *Brugia* adult worm proteins by peptide mass fingerprinting. *J. Parasitol.*, **95**, 1429–1435.

16 Stefanić, S., Dvořák, J., Horn, M., Braschi, S., Sojka, D., Ruelas, D.S., Suzuki, B., Lim, K. *et al.* (2010) RNA interference in *Schistosoma mansoni* schistosomula: selectivity, sensitivity and operation for larger-scale screening. *PLoS Negl. Trop. Dis.*, **4**, e850.

17 Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2011) GenBank. *Nucleic Acids Res.*, **39**, D32–D37.

18 Hertz-Fowler, C., Peacock, C.S., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M. *et al.* (2004) GeneDB: a resource for prokaryotic and

eukaryotic organisms. *Nucleic Acids Res.*, **32**, D339–D343.

19 Aurrecoechea, C., Brestelli, J., Brunk, B.P., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G. *et al.* (2010) EuPathDB: a portal to eukaryotic pathogen databases. *Nucleic Acids Res.*, **38**, D415–D419.

20 Harris, T.W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W.J., De La Cruz, N., Davis, P. *et al.* (2010) WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.*, **38**, D463–D467.

21 Kumar, S., Chaudhary, K., Foster, J.M., Novelli, J.F., Zhang, Y., Wang, S., Spiro, D., Ghedin, E. *et al.* (2007) Mining predicted essential genes of *Brugia malayi* for nematode drug targets. *PLoS One*, **2**, e1189.

22 Caffrey, C.R., Rohwer, A., Oellien, F., Marhöfer, R.J., Braschi, S., Oliveira, G., McKerrow, J.H., and Selzer, P.M. (2009) A comparative chemogenomics strategy to predict potential drug targets in the metazoan pathogen, *Schistosoma mansoni. PLoS One*, **4**, e4413.

23 Chen, F., Mackey, A.J., Stoeckert, C.J., and Roos, D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.

24 Agüero, F., Al-Lazikani, B., Aslett, M., Berriman, M., Buckner, F.S., Campbell, R.K., Carmona, S., Carruthers, I.M. *et al.* (2008) Genomic-scale prioritization of drug targets: the TDR Targets database. *Nat. Rev. Drug Discov.*, **7**, 900–907.

25 *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode, *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.

26 Ghedin, E., Wang, S., Spiro, D., Caler, E., Zhao, Q., Crabtree, J., Allen, J.E., Delcher, A.L. *et al.* (2007) Draft genome of the filarial nematode parasite *Brugia malayi. Science*, **317**, 1756–1760.

27 Berriman, M., Haas, B.J., LoVerde, P.T., Wilson, R.A., Dillon, G.P., Cerqueira, G.C., Mashiyama, S.T., Al-Lazikani, B. *et al.* (2009) The genome of the blood fluke *Schistosoma mansoni. Nature*, **460**, 352–358.

28 Mitreva, M., Jasmer, D.P., Zarlenga, D.S., Wang, Z., Abubucker, S., Martin, J., Taylor, C.M., Yin, Y. *et al.* (2011) The draft genome of the parasitic nematode *Trichinella spiralis. Nat. Genet.*, **43**, 228–235.

29 Martin, J., Abubucker, S., Wylie, T., Yin, Y., Wang, Z., and Mitreva, M. (2009) Nematode.net update 2008: improvements enabling more efficient data mining and comparative nematode genomics. *Nucleic Acids Res.*, **37**, D571–D578.

30 Zerlotini, A., Heiges, M., Wang, H., Moraes, R.L.V., Dominitini, A.J., Ruiz, J.C., Kissinger, J.C., and Oliveira, G. (2009) SchistoDB: a *Schistosoma mansoni* genome resource. *Nucleic Acids Res.*, **37**, D579–D582.

31 Deng, J., Deng, L., Su, S., Zhang, M., Lin, X., Wei, L., Minai, A.A., Hassett, D.J. *et al.* (2011) Investigating the predictability of essential genes across distantly related organisms using an integrative approach. *Nucleic Acids Res.*, **39**, 795–807.

32 McCarter, J.P. (2004) Genomic filtering: an approach to discovering novel antiparasitics. *Trends Parasitol.*, **20**, 462–468.

33 Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.

34 Wang, C.C. (1997) Validating targets for antiparasite chemotherapy. *Parasitology*, **114** (Suppl.), S31–S44.

35 Hardy, L.W. and Peet, N.P. (2004) The multiple orthogonal tools approach to define molecular causation in the validation of druggable targets. *Drug Discov. Today*, **9**, 117–126.

36 Carter, N.S. and Fairlamb, A.H. (1993) Arsenical-resistant trypanosomes lack an unusual adenosine transporter. *Nature*, **361**, 173–176.

37 Maus, C.E., Plikaytis, B.B., and Shinnick, T.M. (2005) Molecular analysis of cross-resistance to capreomycin, kanamycin, amikacin, and viomycin in *Mycobacterium tuberculosis. Antimicrob. Agents Chemother.*, **49**, 3192–3197.

38 Behm, C.A., Bendig, M.M., McCarter, J.P., and Sluder, A.E. (2005) RNAi-based discovery and validation of new drug targets in filarial nematodes. *Trends Parasitol.*, **21**, 97–100.

39 Viney, M.E. (2006) The biology and genomics of *Strongyloides*. *Med. Microbiol. Immunol.*, **195**, 49–54.

40 Doyle, M.A., Gasser, R.B., Woodcroft, B.J., Hall, R.S., and Ralph, S.A. (2010) Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics*, **11**, 222.

41 Batada, N.N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B., Hurst, L.D., and Tyers, M. (2006) Stratus not altocumulus: a new view of the yeast protein interaction network. *PLoS Biol.*, **4**, e317.

42 Jeong, H., Mason, S.P., Barabási, A.L., and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

43 Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A.G., and Marcotte, E.M. (2008) A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nat. Genet.*, **40**, 181–188.

44 Wylie, T., Martin, J., Abubucker, S., Yin, Y., Messina, D., Wang, Z., McCarter, J.P., and Mitreva, M. (2008) NemaPath: online exploration of KEGG-based metabolic pathways for nematodes. *BMC Genomics*, **9**, 525.

45 Crowther, G.J., Shanmugam, D., Carmona, S.J., Doyle, M.A., Hertz-Fowler, C., Berriman, M., Nwaka, S., Ralph, S.A. *et al.* (2010) Identification of attractive drug targets in neglected-disease pathogens using an *in silico* approach. *PLoS Negl. Trop. Dis.*, **4**, e804.

46 Parkinson, J., Whitton, C., Schmid, R., Thomson, M., and Blaxter, M. (2004) NEMBASE: a resource for parasitic nematode ESTs. *Nucleic Acids Res.*, **32**, D427–D430.

47 Geary, T.G., Klein, R.D., Vanover, L., Bowman, J.W., and Thompson, D.P. (1992) The nervous systems of helminths as targets for drugs. *J. Parasitol.*, **78**, 215–230.

48 Chikina, M.D., Huttenhower, C., Murphy, C.T., and Troyanskaya, O.G. (2009) Global prediction of tissue-specific gene expression and context-dependent gene networks in *Caenorhabditis elegans*. *PLoS Comput. Biol.*, **5**, e1000417.

49 Jackson, A.L., Bartz, S.R., Schelter, J., Kobayashi, S.V., Burchard, J., Mao, M., Li, B., Cavet, G. *et al.* (2003) Expression profiling reveals off-target gene regulation by RNAi. *Nat. Biotechnol.*, **21**, 635–637.

50 Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlic, A., Quesada, M. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–401.

51 Magariños, M.P., Carmona, S.J., Crowther, G.J., Ralph, S.A., Roos, D.S., Shanmugam, D., Van Voorhis, W.C., and Agüero, F. (2012) TDR Targets: a chemogenomics resource for neglected diseases. *Nucleic Acids Res.*, **40**, D1118–D1127.