INNOVATION

Genomic-scale prioritization of drug targets: the TDR Targets database

Fernán Agüero, Bissan Al-Lazikani, Martin Aslett, Matthew Berriman, Frederick S. Buckner, Robert K. Campbell, Santiago Carmona, Ian M. Carruthers, A. W. Edith Chan, Feng Chen, Gregory J. Crowther, Maria A. Doyle, Christiane Hertz-Fowler, Andrew L. Hopkins, Gregg McAllister, Solomon Nwaka, John P. Overington, Arnab Pain, Gaia V. Paolini, Ursula Pieper, Stuart A. Ralph, Aaron Riechers, David S. Roos, Andrej Sali, Dhanasekaran Shanmugam, Takashi Suzuki, Wesley C. Van Voorhis and Christophe L. M. J. Verlinde

Abstract | The increasing availability of genomic data for pathogens that cause tropical diseases has created new opportunities for drug discovery and development. However, if the potential of such data is to be fully exploited, the data must be effectively integrated and be easy to interrogate. Here, we discuss the development of the TDR Targets database (http://tdrtargets.org), which encompasses extensive genetic, biochemical and pharmacological data related to tropical disease pathogens, as well as computationally predicted druggability for potential targets and compound desirability information. By allowing the integration and weighting of this information, this database aims to facilitate the identification and prioritization of candidate drug targets for pathogens.

Drug development is urgently needed to combat infectious diseases in the developing world such as malaria, tuberculosis, African trypanosomiasis, Chagas disease, leishmaniasis, onchocerciasis, lymphatic filariasis and schistosomiasis. Even in cases in which drugs for these diseases are available, their use is often limited by factors including high cost, low efficacy, toxicity and the emergence of resistance[1].

Despite this pressing need, the development of new therapeutics to combat these diseases has been inadequate for reasons ranging from a limited understanding of targets that might be amenable to drug development, to anticipated low return on investment. However, recent trends are encouraging[1]. Philanthropic organizations have helped to kindle interest in tropical disease research, and the advent of public–private partnerships has stimulated collaborations between academia and the pharmaceutical industry. On the basic science front, the genome sequences of the disease-causing pathogens are now becoming available, and new technologies are aiding the evaluation of gene function, essentiality and suitability for drug development.

To facilitate the assimilation, integration and mining of data emerging from such studies, and the identification and prioritization of candidate drug targets, we have established a global network of public and private sector partners to develop an open-access database for tropical disease pathogens: the TDR Targets database (http://tdrtargets.org). This resource seeks to bring together data and annotation emerging from genome sequencing and functional genomics projects, protein structural data, manual curation of inhibitors and targets, and information on target essentiality and druggability. We do not propose that this (or any) *in silico* strategy will be able to identify targets for successful drug development through computational methods alone. Rather, our goal is to facilitate the translation of biological questions into a computationally tractable format, enabling individual researchers to query the database, scan the vast quantity of genomic-scale data sets that are now available, and filter out and prioritize a shortlist of candidate targets that are suitable for further investigation.

As of July 2008, the TDR Targets database provides resources for the exploration of drug targets in the tuberculosis pathogen *Mycobacterium tuberculosis*; the leprosy pathogen *Mycobacterium leprae*;

Table 1 | **Selected lines of evidence accessible via the TDR Targets database***

Criteria	*Mycobacterium tuberculosis*	*Plasmodium falciparum*	*Leishmania major*	*Trypanosoma brucei*	*Trypanosoma cruzi*
Enzymes	1,790	862	2,100	1,918	5,474
Assayability	192	205	255	283	252
Crystal structures	229	102	37	80	49
Model structures	2,756	4,829	3,973	7,119	7,103
Phylogeny; not in mammals	3,294	3,850	6,606	8,494	20,730
Essentiality; any evidence	1,054	1,162	1,482	1,375	2,313
Phenotype data; curated	93	107	171	301	In progress
Druggability‡; D index > 0.6	16	127	230	205	325
Compound desirability‡ > 0.6	31	23	37	33	54
Compound links; curated	56	70	129	20	In progress
Compound links; DrugBank	In progress	115	260	287	477
Compound links; PubMed (CAS +EC)	590	351	352	533	435

*As of July 2008. ‡For further information, please see Supplementary information S1 (box). CAS, Chemical Abstracts Service; EC, Enzyme Commission number.

and the malaria parasites *Plasmodium falciparum* and *P. vivax*. The database also includes genome data for the intracellular protozoan parasite *Toxoplasma gondii*; the filariasis helminth pathogen *Brugia malayi* and its intracellular symbiont bacterium *Wolbachia bancrofti*; and the kinetoplastid parasites *Leishmania major*, *Trypanosoma brucei* and *T. cruzi*, which are responsible for kala-azar and other forms of leishmaniasis, sleeping sickness, and Chagas disease, respectively (TABLE 1; see Supplementary information S1 (box)).

Key features of the TDR Targets database include:

- Incorporation of a wide range of genetic, genomic, biochemical, structural and pharmacological data from diverse sources.
- Computational assessment of target druggability and compound desirability.
- Orthology-based inference of relevant information for genes lacking direct functional evidence in particular species of interest.
- Integration of these large-scale data sets with manually curated information on genetic and chemical validation, collected from the primary literature and community surveys.
- The ability to weight results so as to assemble a ranked list of candidate targets. These results may be saved for future reference, modified as new information becomes available, downloaded for integration with locally held data sets, or posted for others to view, modify or download.

To date, virtually all approved anti-infective drugs have been discovered and developed via non-target-based approaches; that is, without optimization for specific targets. Notable exceptions include alpha-difluoromethylornithine, which inhibits ornithine decarboxylase, for African sleeping sickness[2]; HIV protease inhibitors[3]; and zanamivir and oseltamivir, which inhibit the neuraminidase enzyme of influenza virus[4]. The ability to rapidly and effectively locate, capture, integrate, query and retrieve genomic-scale data sets should greatly expedite target-based drug discovery efforts against tropical infectious diseases. In this article, we describe the characteristics of the TDR Targets database and discuss how we have approached the associated challenges for data integration and application.

Box 1 | **Definition of important criteria used in the TDR Targets database**

Orthologues and paralogues. Refer to homologous genes that arise either by speciation from a common ancestor (in the case of orthologues) or by gene duplication (in the case of paralogues). True orthologues tend to be functionally conserved, whereas paralogues, arising from ancient duplication, can be functionally divergent. Orthologue groups are built by clustering orthologous proteins from multiple taxa. Such groupings provide the framework for comparing genes across multiple species and therefore provide information for the functional annotation of genes. For example, in the case of parasitic species, one can gain insights on target selectivity by identifying genes that are either missing or sufficiently divergent from the host species. For more information on orthologue grouping (see OrthoMCL database web site in Further information).

Phylogenetic distribution. Refers to the presence or absence of orthologous genes in a chosen group of organisms. Information regarding gene distribution is obtained from orthologue identification and grouping (see above).

Essentiality. Term used to describe genes that are essential for the growth and survival of an organism. Information on gene essentiality for an organism is gathered mainly from genomic-scale experimental data sets. When such data sets are lacking, especially for parasitic species, data available for orthologous genes in model organisms are mapped to the corresponding parasite genes, thereby suggesting precedence for essentiality.

Assayability. Refers to the feasibility of performing an assay for an enzyme based on the availability of protocols and reagents. For example, whether there is precedent for readout of target activity using a biochemical assay or precedent for production of recombinant protein in soluble form. However, no effort is made to assess the ease of assay implementation.

Druggability. Provides a measure of whether a gene of interest from a parasitic organism can be targeted by compounds that are likely to be efficacious in and tolerated by the host organism.

Compound desirabilty. Provides a measure of the chemical quality of inhibitors that target a gene (or most similar gene) of interest. Compounds that have good inhibitory effects but have toxic and reactive side groups that are essential for inhibition are not considered as ideal leads for drug development.

Curated annotation. Refers to the manual curation effort by members of the TDR Drug Targets network for the purpose of collecting and storing (as structured ontology) information regarding the observed phenotypic effect on the parasite that is due to either a genetic or a chemical perturbation. This also includes data from community-wide surveys on potential targets.

For more information on these criteria see the methods section in Supplementary information S1 (box).

Database content

The potential of a given gene product of a pathogen as a therapeutic target depends on two broad types of information. First, the role of the gene in the pathogen and second, the likelihood of being able to develop a compound that targets that gene product. The design and functionality of the database allows users to address both of these criteria.

The role of potential targets. The database is structured to incorporate genome sequence information as well as functional genomics and available essentiality data (BOX 1; see Supplementary information S1 (box)). In less than a decade, high-quality reference genome sequences have become available for the major parasite groups, including *Babesia*, *Cryptosporidium*, *Giardia*, *Entamoeba*, *Leishmania*, *Plasmodium*, *Trichomonas*, and the African and American trypanosomes. Until recently, pathogenic worms were a striking exception, but the genome sequence of the filarial *Brugia* worm has recently been published[5], and at least a dozen helminth genome projects are in progress. Comparison of sequences from multiple related *Plasmodium*, *Leishmania* and *Trypanosoma* species improves the accuracy of gene finding and provides an opportunity to identify species-specific genes and signals of evolutionary selective pressure that could potentially aid target prioritization. A key part of the database development is therefore to work closely with those who produce sequence data to

“The ability to rapidly and effectively locate, capture, integrate, query and retrieve genomic-scale data sets should greatly expedite target-based drug discovery efforts against tropical infectious diseases.”

include new genomes as they become available, and to streamline data loading so that information can be updated from files provided by external sources.

Although functional genomic data are becoming available for many pathogens of interest, technological limitations currently preclude genome-wide analysis for all species considered here except *M. tuberculosis*[6–9]. Thanks to manual curation of the literature, the TDR Targets database includes information on the roles of individual genes for which phenotype data are published. Where information is not directly available for organisms included in the database, insight may sometimes be inferred through orthology. Based on an evaluation of available orthologue identification methodology[10], we have used OrthoMCL[11] (see Further information) to map all genes from targeted pathogens to orthologues in organisms for which additional information is available. High-quality whole-genome inactivation data are available for *Saccharomyces cerevisiae* (in the form of genetic lesions[12–14]) and *Caenorhabditis elegans* (in the form of RNA interference knockdowns[15]), and essentiality data on orthologues in bacteria[16] may be relevant especially for bacterial pathogens and endosymbiotic organelles in eukaryotes. Inferring essentiality by orthology can be risky — functional redundancy in a pathogen of interest may yield false positives (for example, dihydrofolate reductase is essential in most species, but not in *L. major*[17]) and false negatives (for example, hypoxanthine phosphoribosyl transferase activity is dispensable in most species, but not in *P. falciparum*[14,18]). However, extensive experience with such analyses suggests that 'essential in at least one organism' is a useful criterion for drug-target prioritization as such genes are more likely to be attractive targets for drug development than those that are not essential in any known organism.

The TDR Targets database therefore allows users to assess the role of the gene in the biology of the pathogen and to predict whether pharmacological targeting of this role is likely to kill the pathogen. Knowledge of orthology, particularly of whether the gene product lacks an orthologue in humans, is also crucial for minimizing the potential for adverse events.

Potential for drug development. The second type of information in the database relates to the likelihood of being able to develop a drug-like compound to modulate the target. For example, structural information for the target or related proteins could aid drug design, and if inhibitors for the target in question or for related proteins are already available, these might provide useful starting points for lead discovery. More generally, the druggability[19] of a given target in a pathogen can be predicted on the basis of factors such as the physicochemical nature

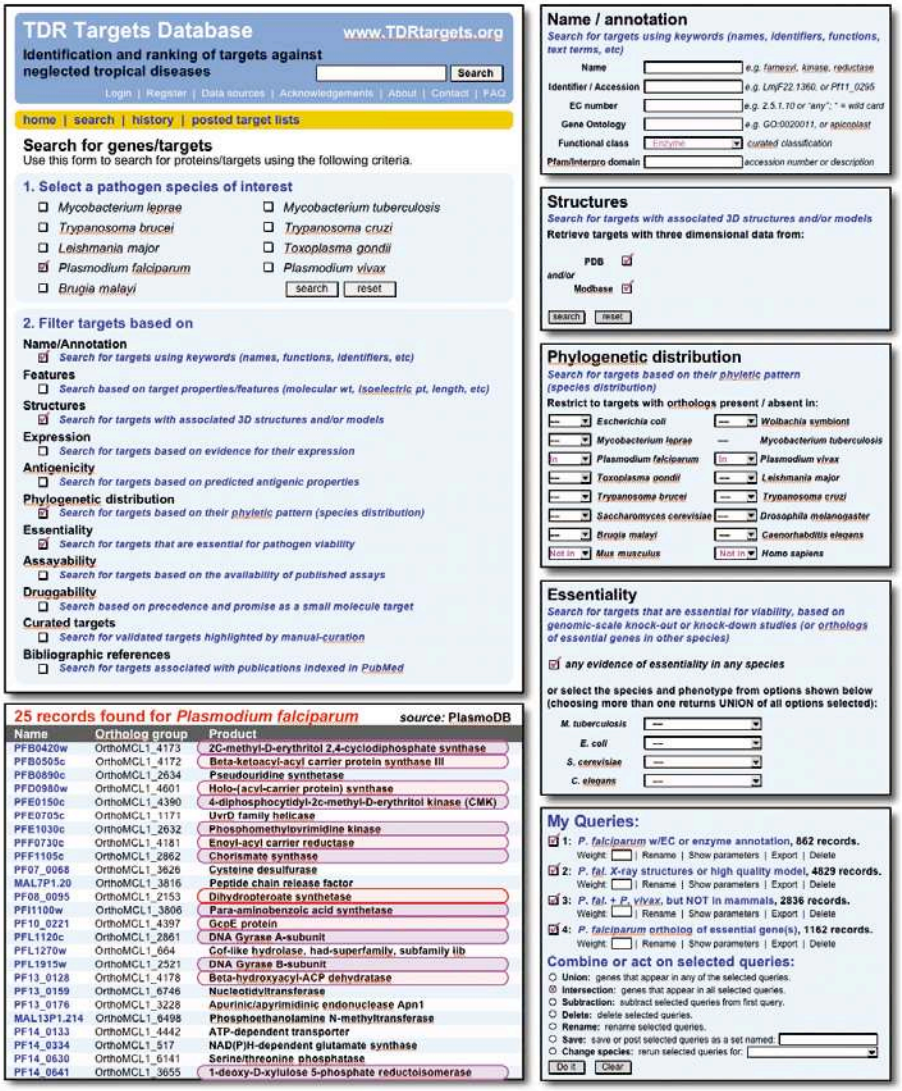

Figure 1 | **Searching the TDR Targets database.** This figure shows the appearance of the search page and how different searches can be combined to yield lists of potential drug targets. First, the search can be restricted to a given species or groups of species; in this case *Plasmodium falciparum* is selected (top panel left). Next, one can query on function-based information. For example, selecting Enzyme in under the Functional class category of the Name/Annotation panel (top right) leads to a list of 862 genes identified in *P. falciparum* as encoding enzymes (bottom right). Similarly, the Structures panel below allows identification of 4,829 genes for which there are crystal structures and/or high-quality structural models from the Phylogenetic distribution panel (middle right). Searching for orthologues present in *Plasmodium vivax* but not in humans or mice leads to 2,836 genes conserved in *Plasmodium* species and absent in mammals. Very few genes have been validated as essential in *P. falciparum*, but an Essentiality search shows 1,162 *P. falciparum* genes with essential orthologues in at least one model organism — *Caenorhabditis elegans*, *Escherichia coli*, *Mycobacterium tuberculosis* and *Saccharomyces cerevisiae* — and are therefore plausibly essential in *P. falciparum*. Each of these independent searches is stored in the History page of the database (bottom right). Registered users' queries are automatically saved for later sessions. The intersection of all four queries shown here have identified 25 potential drug targets for *P. falciparum*. These *P. falciparum* genes are potentially selective drug targets (they lack orthologous targets in mammals) that might allow for structure-based design of drugs (using available structures or models), are probably assayable (all are enzymes) and might be essential for viability (based on an orthologue being essential in some model organism). There are many promising drug targets among this list (circled in red), supporting this approach as one means to prioritize promising targets.

of small-molecule binding sites on the target and the availability of drug-like molecules that target related proteins in other organisms. The ease of developing an assay to screen for compounds that modulate the activity of the target protein is also an important consideration.

Structural biology, and the potential for structure-guided drug design, is another area in which high-quality data from model organisms can be translated to under-studied pathogens based on orthology mapping. Only a small number of pathogen proteins have been structurally characterized to date, although initiatives to remedy this include projects by the Medical Structural Genomics of Pathogenic Protozoa (MSGPP) and the Structural Genomics Consortium (see Further information). Nevertheless, high-confidence structural models have been generated for many parasite proteins based on sequence similarity to orthologous proteins for which crystal structures are available[20]. Gene pages in the TDR Targets database link to these models at ModBase (see Further information). Potential inhibitors may also be inferred through orthology to proteins for which crystal structures or models of inhibitor–ligand interactions are available.

The TDR Targets database predicts the druggability of each parasite protein using various methods (see Supplementary information S1 (box)). Precedence for the druggability of targets was derived from Pfizer's comprehensive survey of known biological targets of drugs, leads and chemical tools[21] and from Biofocus's StARlite database (which is scheduled to become publicly available through the European Bioinformatics Institute (EBI); see Further information) of the medicinal chemistry literature. The sequences of approximately 1,400 proteins with known drug-like ligands were mapped onto pathogen genomes using a high-confidence orthology method[10,11] and a lower-confidence sequence homology (basic local alignment tool; BLAST) method to identify sets of orthologous and homologous druggable targets. Druggability was also inferred using a sequence-feature-based Bayesian algorithm trained on a set of known drug targets[12]. The druggability of each of the protein models generated for each parasite in the database was also assessed using a structure-based binding-site algorithm[13]. A normalized, weighted sum based on the accumulation of prediction for the different methods results in a composite Druggability Confidence Index, with a value ranging between 0 and 1 (with 1 as the ideal value) for each parasite protein. This index also reflects the degree of similarity between the pathogen target and the known druggable homologue. A Compound Desirability Index (also ranging between 0 to 1) was also assigned to each target gene based on the chemical quality of known inhibitors of the most similar target in other species (see Supplementary information S1 (box)). Although the database focuses on the identification and prioritization of drug targets from a gene-centric and protein-centric perspective, links to chemical compounds are also included; this provides the ability to search for genes that are targeted by specific compounds (see Supplementary information S1 (box)).

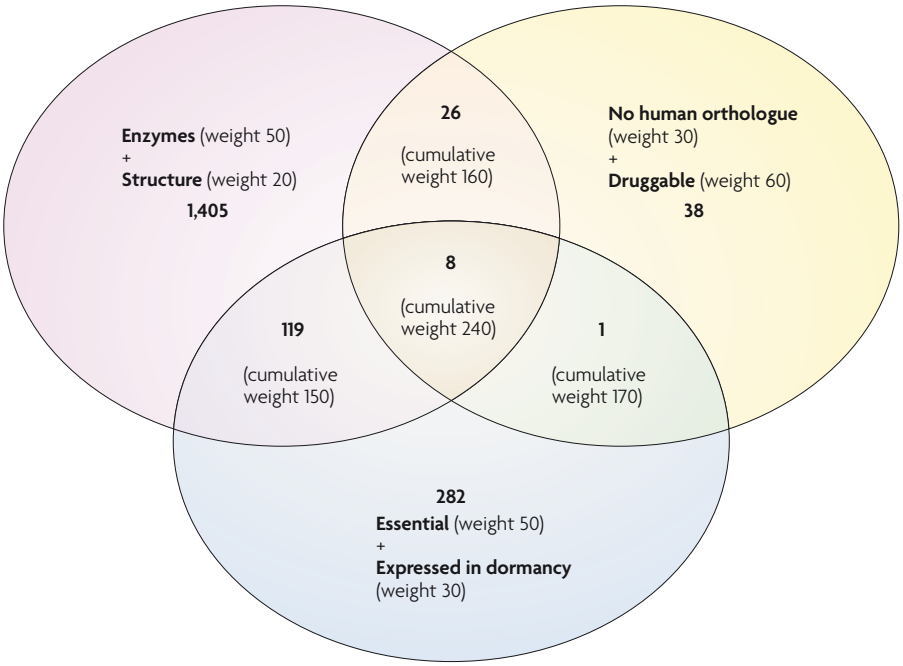

Figure 2 | **Ranking of *Mycobacterium tuberculosis* targets using the TDR Targets database.** This Venn diagram demonstrates the gene-ranking functionality of the TDR Targets database. This example shows the weighted union of queries to yield a list of ranked genes from *Mycobacterium tuberculosis*. The selected criteria, the arbitrary weights assigned to each criteria and the total number of genes returned for each query (bold font) are indicated in the diagram. Cumulative weights were assigned to genes intersecting multiple queries by summing the total weight of individual queries. For instance, the eight genes that intersect all three queries — Query 1: enzymes (weighted 50) + having structure or structure model (weighted 20); Query 2: no human orthologue (weighted 30) + druggable (weighted 60); Query 3: essential in any species (weighted 50) + expressed in dormancy (weighted 30) — get the highest cumulative weight of 240. TABLE 2 shows a subset of the 1,563 total genes selected by the above criteria combination, showing the eight highest ranked genes in the top left column. Many of these highly ranked genes code for cytochrome P450 enzymes, which have homologues in humans. However, they are selected in Query 2 (as well as the other queries) because they are sufficiently distinct from their human counterparts that they cluster into different orthologue groups. This suggests potential target selectivity.

Information on the availability of enzyme assays and reagents (assayability) is also useful for the purpose of target validation. The database provides assayability information for genes where available. In cases for which an Enzyme Commission (EC) number exists for a target protein, a link is provided to a protocol for that enzyme assay (generally based on a model organism) in the Sigma–Aldrich Enzyme Explorer Assay library (see Further information). A total of 1,707 genes from the target organisms currently have such links to published assays. Development of an enzyme assay for many of these organisms is often hindered by difficulties in producing soluble recombinant protein. Information on precedence for producing soluble recombinant protein is available in the curated literature and from consortia such as the MSGPP that generate such reagents for structural purposes in high-throughput efforts. Based on data available from the MSGPP, 702 recombinant proteins or fragments have been successfully produced and mapped to the corresponding pathogen genes in TDR targets.

Table 2 | **Ranked list of *Mycobacterium tuberculosis* targets: top 50 out of total 1,563 records**

Gene ID	Product name	Weight	Gene ID	Product name	Weight
Rv0766c	Cytochrome P450 Cyp123	240	Rv0129c	Secreted antigen 85 complex C antigen (FbpC)	160
Rv1256c	Cytochrome P450 Cyp130	240	Rv1338	Glutamate racemase (MurI)	160
Rv2191	Conserved hypothetical protein	240	Rv1523	Methyltransferase	160
Rv2266	Cytochrome P450 Cyp124	240	Rv0769	Dehydrogenase/reductase	160
Rv3545c	Cytochrome P450 Cyp125	240	Rv1599	Histidinol dehydrogenase (HisD)	160
Rv1880c	Cytochrome P450 Cyp140	240	Rv1484	NADH-dependent Enoyl-ACP Reductase (InhA)	160
Rv3518c	Cytochrome P450 Cyp142	240	Rv1547	DNA polymerase III, α-subunit (DnaE1)	160
Rv1284	Conserved hypothetical protein	240	Rv1601	Imidazole glycerol-phosphate dehydratase (HisB)	160
Rv1730c	Possible penicillin-binding protein	170	Rv2268c	Cytochrome P450 Cyp128	160
Rv0533c	β-Ketoacyl-ACP synthase III (FABH)	160	Rv2157c	UDP-*N*-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate-D-alanyl-D-alanyl ligase (MurF)	160
Rv2870c	DOX-P reductoisomerase	160	Rv1131	Citrate synthase I (GltA1)	150
Rv1886c	Secreted antigen 85 complex B (FbpB)	160	Rv1018c	UDP-*N*-Ac-glucosamine pyrophosphorylase (GlmU)	150
Rv0091	Bifunctional Mta/Sah nucleosidase	160	Rv0764c	Cytochrome P450 Cyp51	150
Rv2156c	P-*N*-acetylmuramoyl-pentapeptidetransferase (MurX)	160	Rv0886	NADPH:adrenodoxin oxidoreductase (FprB)	150
Rv295 2	Methyltransferase	160	Rv1348	ABC transporter	150
Rv3711c	DNA polymerase III, ε-subunit (DnaQ)	160	Rv1714	Oxidoreductase	150
Rv3804c	Secreted antigen 85 complex A (FbpA)	160	Rv2122c	Phosphoribosyl-AMP pyrophosphatase (HisE)	150
Rv2068c	Class A β-lactamase	160	Rv2794c	Conserved hypothetical protein	150
Rv2537c	3-Dehydroquinate dehydratase Arod (AroQ)	160	Rv0600c	Two component sensor kinase	150
Rv2538c	3-Dehydroquinate synthase (AroB)	160	Rv2964	Formyltetrahydrofolate deformylase (PurU)	150
Rv2523c	Acyl carrier protein synthase	160	Rv0670	Endonuclease IV	150
Rv0778	Cytochrome P450 Cyp126	160	Rv1463	ABC transporter	150
Rv2540c	Chorismate synthase (Arof)	160	Rv2220	Glutamine synthetase (GlnA1)	150
Rv3803c	Secreted Mpt51/Mpb51 antigen 85 complex C (FbpD)	160	Rv0983	Serine protease (PepD)	150
Rv3227	3-P-shikimate 1-carboxyvinyltransferase (AroA)	160	Rv3427c	Possible transposase	150

Querying the database

We envisage that most users will initiate a data-mining session with a preconceived set of attributes that they deem desirable or necessary for drug-target selection. These attributes may relate to the essentiality of the target (for example, determined by genetic and/or pharmacological validation or inferred from metabolic pathway maps); the suitability of the target for expression and assayability; the availability of structures or models with which to initiate rational drug design; the precedence for druggability; and/or potential for inhibitor selectivity. Preferences will differ among individuals and organizations — depending on experience with particular target classes and assay systems — the availability and diversity of compound libraries, and strategies for hit identification. Furthermore, different organisms will probably require different search strategies, due in part to differences in the availability and usefulness of specific data sets for each organism. For example, the paucity of genetic validation data for apicomplexan parasites[22] may require users to infer the essentiality of their genes indirectly via other criteria. The TDR Targets database creates a structure within which a wide range of queries can be articulated, while prompting users to define parameters for potentially important criteria that they may not have considered previously. Step-by-step examples of such searches are shown in FIG. 1 and in Supplementary information S2 (figure); all search results shown are based on data available as of July 2008.

Compilation of diverse data types that are relevant to drug-target discovery in a single location is in itself a valuable undertaking. Incorporation of this information into a relational database, however, allows the formulation of complex queries; for example, "find all *T. cruzi* genes that are predicted to encode essential enzymes but that are absent from the human host". The TDR Targets database supports Boolean operations (union/OR, intersection/AND), which allows complex queries such as the above example to be assembled from a series of simpler queries. In particular, the intersection of queries is a powerful tool for reducing a large number of entries to a manageable list that might be prioritized for further experimental studies. It is also particularly useful when a user has several absolute conditions that acceptable targets must meet to be further considered. Intersecting queries in the TDR Targets database can be accomplished by choosing multiple features on the primary search page (see FIG. 1 for an example), or can be generated by combining multiple queries

Box 2 | **Generating a ranked list of putative drug targets in *Trypanosoma brucei****

In this example — table on the right shows the top 24 target genes, ranked by weight, of 10,965 genes — the highest-scoring target for *Trypanosoma brucei* is farnesyl pyrophosphate synthase, a protein that additional experimental work has suggested is a promising drug target[26]. Clicking on the name of the target (Tb927.7.3360) in the web site leads to a target-specific page showing the following information:

- This target has been genetically validated by RNA interference to demonstrate a growth deficit in the bloodstream (mammalian stage).
- It is a 42 kDa enzyme and its orthologues in *Caenorhabditis elegans* and *Saccharomyces cerevisiae* are essential.
- It has a druggability score of 0.8 (in which 1 is optimal druggability) and a compound desirability score of 0.3 (in a range of 0–1 based on the interactions of 97 inhibitors with orthologues).
- There are literature links to 12 interacting chemical compounds with this enzyme in *T. brucei*.
- There are two structures for this enzyme in Protein Data Bank as well as a ModBase model.
- The database has curated 12 bibliographical references for the enzyme.
- Genetic and chemical validation experiments have been published on this enzyme in *T. brucei*.

All of the other genes in the ranked results can be similarly examined in depth by clicking on the target names. This list can be exported into a tab-delimited file and manipulated as a spreadsheet. Please see Supplementary information S2 (figure) for a pictorial summary of this multiple-query search. Search criteria and weights (in brackets) are outlined below (maximum possible cumulative weight is 465):

- Enzymes (100)
- Low mass <100 kDa (20)
- No transmembrane domains (20)
- Crystal structure (50)
- Structural model in Modbase (30)
- Present in all trypanosomatids (25)
- Absent in humans (25)
- Essential in at least one model organism (40)
- Druggability > 0.6 (35)
- Compound desirability > 0.3 (35)
- Chemical and/or genetic validation (50)
- Publication(s) in PubMed (35)
- Maximum possible cumulative weight (465)

*Data obtained from *T. brucei* query set (DSR VI/11/07); see Further information.

Product name	Gene ID	Weight
Farnesyl pyrophosphate synthase	Tb927.7.3360	405
Phosphoglycerate kinase	Tb927.1.700	370
Fructose-bisphosphate aldolase, glycosomal	Tb10.70.1370	370
Enolase	Tb10.70.4740	370
Protein kinase, putative	Tb927.6.2030	365
UDP-galactose 4-epimerase	Tb11.02.0330	365
Ornithine decarboxylase	Tb11.01.5300	365
Cysteine peptidase C, cathepsin B-like	Tb927.6.560	355
N-Myristoyl transferase, putative	Tb10.61.2550	355
Prostaglandin F synthase	Tb11.02.2310	355
Dihydrofolate reductase-thymidylate synthase	Tb927.7.5480	355
DNA topoisomerase II	Tb09.160.4090	345
ATP synthase F1, beta subunit	Tb927.3.1380	340
CDC2-related protein kinase	Tb10.70.2210	340
Glycogen synthase kinase, putative	Tb10.61.3140	340
Dual-specificity protein kinase, putative	Tb11.02.0640	340
Protein kinase, putative	Tb927.8.5950	340
Receptor-type adenylate cyclase GRESAG 4	Tb11.03.0970	340
Protein kinase, putative	Tb09.160.0570	340
CDC2-like protein kinase	Tb10.70.7040	340
Protein kinase, putative	Tb927.7.6220	340
Protein kinase A catalytic subunit	Tb10.389.0490	340
V-type ATPase, A subunit, putative	Tb927.4.1080	340
Pteridine reductase, putative	Tb927.8.2210	340

in the user's history page, which saves every query for registered users and can be revisited in later sessions. Once the list is generated, individual pages for each gene can be followed for further information on known inhibitors, notes on genetic or chemical validation and other manually curated data. The list, with customized data types, can also be exported into a spreadsheet or tab-delimited file and manipulated using spreadsheet applications.

Although intersecting queries can be a powerful method for reducing the number of potential targets to those of particular interest, these queries will eliminate targets that should be considered but which fail to meet only one or a few of the specified criteria. The incompleteness of available experimental data means that in many cases intersection queries may be too stringent for prioritizing drug targets. The TDR Targets database allows the user to combine a union of queries (Boolean OR) and to subsequently rank the union to generate an ordered list of all targets that meet at least one criterion. To generate a ranking according to the user's preferences, individual queries are each assigned a numerical weight by the user and then combined, with the end list ranked according to the additive weighting. The user assigns higher numbers to features considered to be particularly important and lower numbers to features that are desirable but not indispensable. FIGURE 2 shows this in a Venn diagram, and the corresponding ranked list is shown in TABLE 2. Using multiple criteria, each weighted differently, the highest ranked targets are found in the intersection and therefore head the list. The example in BOX 2 and in Supplementary information S2 (figure) shows how the user might combine multiple queries to return a ranked list, which are headed by genes that fulfil most of that user's highly ranked criteria, followed by genes that fulfil gradually

fewer of the nominated criteria. Weighting values that produce suboptimal or biased rankings can be adjusted on the history page and reapplied to iteratively improve the usefulness of the prioritization process.

A key feature of the database is that it allows users to view and share the results of their analyses on a community-wide basis. This allows users to integrate query results that have been generated by other users into their own analysis. Saved query sets can be posted (or published) by using the "publish" functionality that is shown under each saved query on the "history" page. Users are prompted to enter a description of the query sets they are posting. All posted data can be viewed from the "posted lists of targets" page. By clicking on the query set names, users can view a list of individual queries and a description of the queries under that set. From this page, one can also view the query results by clicking on the individual query names. Users can import queries by selecting the relevant queries and clicking on the "import into my history" button. Once imported, the queries will appear on the user's history page and are then available for combining with any of the other queries listed therein. Any data posted by the user will also be listed under the "my published query sets" section of their history. Posted data can be removed by clicking on the relevant "unpublish" button listed under each query.

The ease of sharing queries should allow users to build on previously published prioritization analyses. Published lists of promising targets for *M. tuberculosis*[23] and *Brugia* species[24] have been imported into the TDR Targets web site and then posted for general use (see examples, tuberculosis target prioritization by Hasan *et al.* and *Brugia* targets ranked by Kumar *et al.*; Further information).

Challenges and future perspectives

The database was launched in March 2007 and as of July 2008 has attracted more than 10,000 visits from all over the world, with over 30% of the visits originating from developing countries and/or territories in which the targeted diseases are endemic. In response to comments from users, we have improved functionalities in various ways. For example, we have generated a manually curated list of enzymes (accessible via the Functional class category on the Name/Annotation panel on the search page) because searching for genes with 'any' EC number excludes enzymes for which EC numbers have not been assigned. Several funding proposals to further validate drug targets have cited results from the TDR Targets database as starting points.

Future releases of the database will mirror changes in parasite genome sequencing and will help to identify and prioritize targets for applications in other areas such as diagnostics. Most notably, major international genome sequencing programmes are being developed for parasitic worms. A draft of the genome of *Schistosoma mansoni* should soon be available (M.B. *et al.*, manuscript in preparation), and the genome of *Onchocerca volvulus* and several related nematode genomes (for example, *Schistosoma japonicum*) should be available within the next few years. In the meantime, we plan to add expressed sequence tag information for these nematodes to allow searching for drug targets before the entire helminth genomes are constructed and annotated.

To aid in the diagnosis of an infection that is caused by organisms listed in the TDR Targets database, we have recently added functionality to predict epitopes, and we plan to add other features that should help in the search for new diagnostics.

Future implementations of the TDR Targets database will also integrate curated data from the Braunschweig enzyme database (BRENDA; see Further information). This database includes valuable literature-derived information on organism-specific assays and on the production of recombinant proteins (and relevant clones).

The development and enrichment of this database would not have been possible without the determination and commitment of all members of the Drug Target Prioritization group assembled by the TDR Drug Targets network. Network members regularly exchange information and ideas, and two face-to-face meetings are held each year to discuss progress and future directions of the project, and to publicly demonstrate the use of the database. External experts from industry and academia, as well as select TDR advisory committee members, participate in these meetings, and industrial partners have contributed valuable data and expertise freely. Lessons learned from this work will be useful in optimizing and scaling-up drug discovery networks[25] for both *in silico* and experimental analyses.

A major future challenge is to secure resources for the continued curation and further optimization of the database. Such optimization is likely to require adaptation of the database to accommodate improved models for predicting druggability, gene-disruption data sets, the challenges of prioritizing drug targets in multicellular versus unicellular target organisms, and the rapidly growing availability of medium- and high-throughput compound screening data based on single-target assays, life–death assays and high-content screens. These emerging fields present challenges as such massive data sets will only attain maximum value if they are carefully organized, integrated and annotated, and queryable in user-friendly environments.

Fernán Agüero and Santiago Carmona are at the Instituto de Investigaciones Biotecnológicas, Universidad Nacional de General San Martín, San Martín 1650, Buenos Aires, Argentina.

Bissan Al-Lazikani, Ian M. Carruthers, A. W. Edith Chan and John P. Overington are at BioFocus DPI, 4th Floor Commonwealth House, 1 New Oxford Street, London WC1A 1NU, UK.

Martin Aslett, Matthew Berriman, Christiane Hertz-Fowler, Arnab Pain and Takashi Suzuki are at the Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK.

Frederick S. Buckner, Gregory J. Crowther, Aaron Riechers and Wesley C. Van Voorhis are at the Department of Medicine, Division of Allergy and Infectious Diseases, University of Washington, Seattle, Washington 98195, USA.

Robert K. Campbell is at the Marine Biological Laboratory, Woods Hole, Massachusetts 02543, USA.*

Feng Chen, David S. Roos and Dhanasekaran Shanmugam are at the Department of Biology, and Penn Genomics Institute, University of Pennsylvania, 415 South University Avenue, Philadelphia 19104-6018, USA.

Maria A. Doyle and Stuart A. Ralph are at the Department of Biochemistry & Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Victoria 3010, Australia

Andrew L. Hopkins‡ and Gaia V. Paolini are at Pfizer Global Research and Development, Sandwich, Kent CT13 9NJ, UK.

Gregg McAllister is at Brandeis University, 415 South Street, Waltham, Massachusetts 02453, USA.

Solomon Nwaka is at the World Health Organization/ Special Programme for Research and Training in Tropical Diseases, 1211 Geneva 27, Switzerland.

Ursula Pieper and Andrej Sali are at the Department of Biopharmaceutical Sciences, University of California San Francisco, 1700 4th Street, QB3, 5 South 503B, San Francisco, California 94143-2552, USA.

Christophe L. M. J. Verlinde is at the Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA.

**R.K.C. is also at Pfizer Global Research and Development, Sandwich, UK.*

‡Current address: Division of Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dow Street, Dundee, DD1 5EH, UK.

Correspondence to F.A., M.B., S.N., S.A.R., D.S.R. & W.C.V.V.

e-mails: fernan@unsam.edu.ar; mb4@sanger.ac.uk; NwakaS@who.int; saralph@unimelb.edu.au; droos@sas.upenn.edu; wesley@u.washington.edu

doi:10.1038/nrd2684
Published online 17 October 2008

1. Nwaka, S. & Hudson, A. Innovative lead discovery strategies for tropical diseases. *Nature Rev. Drug Discov.* **5**, 941–955 (2006).
2. Heby, O., Persson, L. & Rentala, M. Targeting the polyamine biosynthetic enzymes: a promising approach to therapy of African sleeping sickness, Chagas' disease, and leishmaniasis. *Amino Acids* **33**, 359–366 (2007).
3. Mori, M. *et al.* Contribution of structural biology to clinically validated target proteins. *Drug Discov. Today* **13**, 469–472 (2008).
4. Varghese, J. N. Development of neuraminidase inhibitors as anti-influenza virus drugs. *Drug Dev. Res.* **46**, 176–196 (1999).
5. Ghedin, E. *et al.* Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317**, 1756–1760 (2007).
6. McAdam, R. A. *et al.* Characterization of a *Mycobacterium tuberculosis* H37Rv transposon library reveals insertions in 351 ORFs and mutants with altered virulence. *Microbiology* **148**, 2975–2986 (2002).
7. Sassetti, C. M., Boyd, D. H. & Rubin, E. J. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* **48**, 77–84 (2003).
8. Lamichhane, G. *et al.* A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **100**, 7213–7218 (2003).
9. McNeil, L. K. *et al.* The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation. *Nucleic Acids Res.* **35**, D347–D353 (2007).
10. Chen, F., Mackey, A. J., Vermunt, J. K. & Roos, D. S. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE* **2**, e383 (2007).
11. Chen, F., Mackey, A. J., Stoeckert, C. J. Jr & Roos, D. S. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **34**, D363–D368 (2006).
12. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nature Rev. Drug Discov.* **5**, 993–996 (2006).
13. Al-Lazikani, B. *et al.* in *Bioinformatics — From Genomes to Therapies. Volume 3: The Holy Grail: Molecular Function* (ed. Lengauer, T.) 1315–1334 (Wiley-VCH, Weinheim, 2007).
14. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
15. Kamath, R. S. *et al.* Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237 (2003).
16. Gerdes, S. Y. *et al.* Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684 (2003).
17. Titus, R. G., Gueiros-Filho, F. J., de Freitas, L. A. & Beverley, S. M. Development of a safe live *Leishmania* vaccine line by gene replacement. *Proc. Natl Acad. Sci. USA* **92**, 10267–10271 (1995).
18. Chaudhary, K. *et al.* Purine salvage pathways in the apicomplexan parasite *Toxoplasma gondii*. *J. Biol. Chem.* **279**, 31221–31227 (2004).
19. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Rev. Drug Discov.* **1**, 727–730 (2002).
20. Pieper, U. *et al.* MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* **32**, D217–D222 (2004).
21. Paolini, G. V., Shapland, R. H., van Hoorn, W. P., Mason, J. S. & Hopkins, A. L. Global mapping of pharmacological space. *Nature Biotech.* **24**, 805–815 (2006).
22. Meissner, M., Breinich, M. S., Gilson, P. R. & Crabb, B. S. Molecular genetic tools in *Toxoplasma* and *Plasmodium*: achievements and future needs. *Curr. Opin. Microbiol.* **10**, 349–356 (2007).
23. Hasan, S., Daugelat, S., Rao, P. S. & Schreiber, M. Prioritizing genomic drug targets in pathogens: application to *Mycobacterium tuberculosis*. *PLoS Comput. Biol.* **2**, e61 (2006).
24. Kumar, S. *et al.* Mining predicted essential genes of *Brugia malayi* for nematode drug targets. *PLoS ONE* **2**, e1189 (2007).
25. Hopkins, A. L., Witty, M. J. & Nwaka, S. Mission possible. *Nature* **449**, 166–169 (2007).
26. Montalvetti, A. *et al.* Farnesyl pyrophosphate synthase is an essential enzyme in *Trypanosoma brucei*. *In vitro* RNA interference and *in vivo* inhibition studies. *J. Biol. Chem.* **278**, 17075–17083 (2003).

Acknowledgements

The authors wish to acknowledge all of the investigators who provided the data in the TDR Targets database including those that participated in the survey on drug targets for Human African Trypanosomiasis (HAT survey) conducted during 2007. We would also like to acknowledge Brandeis University MS students P. Bais and B. Coflan for work on the association of targets with compounds; R. L. Stevens (Argonne National Laboratory) for providing data for gene essentiality in bacteria; K. Chaudhary and T. Carlow (New England BioLabs) for integrated *C. elegans* phenotype data; J. Sacchetini (Texas A&M) for information on known *M. tuberculosis* drug targets; and M. Schreiber (Novartis Institute for Tropical Diseases, Singapore) and J. Brown (GlaxoSmithKline) for input on integrating data on persistent expressed genes in dormant-stage *M. tuberculosis* infection. We would also like to acknowledge essential computational infrastructure and genome annotations made available through the OrthoMCL database (supported by the US National Institutes of Health; NIH); GeneDB (supported by the Wellcome Trust); Ensembl (supported by the European Bioinformatics Institute); and EuPathDB (supported by a Bioinformatics Resource Center contract from the US NIH/National Institute of Allergy and Infectious Diseases). The authors also gratefully acknowledge Pfizer Global Research and Development for sharing data related to druggability. This work was supported by grants from the United Nations Development Programme/World Bank/World Health Organization Special Programme for Research and Training in Tropical Diseases.

FURTHER INFORMATION

BRENDA: http://www.brenda-enzymes.info
***Brugia* targets ranked by Kumar *et al.*:** http://www.tdrtargets.org/published/browse/230
EBI Chemigenomics databases: www.ebi.ac.uk/chembl
Medical Structural Genomics of Pathogenic Protozoa: http://www.msgpp.org
ModBase: http://modbase.compbio.ucsf.edu
OrthoMCL database: http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi
Sigma–Aldrich Enzyme Explorer Assay library: http://www.sigmaaldrich.com/Area_of_Interest/Biochemicals/Enzyme_Explorer/Key_Resources/Assay_Library.html
Structural Genomics Consortium: http://www.sgc.utoronto.ca
***T. brucei* query set (DSR VI/11/07):** http://tdrtargets.org/published/browse/91
TDR Targets database: http://tdrtargets.org
Tuberculosis target prioritization by Hasan *et al.*: http://www.tdrtargets.org/published/browse/226

SUPPLEMENTARY INFORMATION

See online article: S1 (box) | S2 (figure)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF