Published in final edited form as:

Nat Rev Drug Discov. 2008 November ; 7(11): 900–907. doi:10.1038/nrd2684.

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

Fernán Agüero^{1,*}, Bissan Al-Lazikani², Martin Aslett³, Matthew Berriman^{3,*}, Frederick S. Buckner⁴, Robert K. Campbell^{5,8}, Santiago Carmona¹, Ian M. Carruthers², A. W. Edith Chan², Feng Chen⁶, Gregory J. Crowther⁴, Maria A. Doyle⁷, Christiane Hertz-Fowler³, Andrew L. Hopkins^{8,12}, Gregg McAllister⁹, Solomon Nwaka^{10,*}, John P. Overington², Arnab Pain³, Gaia V. Paolini⁸, Ursula Pieper¹¹, Stuart A. Ralph^{7,*}, Aaron Riechers⁴, David S. Roos^{6,*}, Andrej Sali¹¹, Dhanasekaran Shanmugam⁶, Takashi Suzuki³, Wesley C. Van Voorhis^{4,*}, and Christophe L. M. J. Verlinde⁴

¹Universidad Nacional de General San Martín, Buenos Aires (Argentina) ²BioFocus DPI, London (UK) ³Sanger Institute, Hinxton, Cambridgeshire (UK) ⁴University of Washington, Seattle WA (USA) ⁵Marine Biological Laboratory, Woods Hole MA (USA) ⁶University of Pennsylvania, Philadelphia PA (USA) ⁷University of Melbourne, Victoria (Australia) ⁸Pfizer Laboratories, Sandwich (UK) ⁹Brandeis University, Waltham MA (USA) ¹⁰WHO/TDR, Geneva (Switzerland) ¹¹University of California, San Francisco CA (USA)

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 TDRtargets.org 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 prioritisation 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 where drugs for such

*Corresponding Authors: Agüero, fernan@unsam.edu.ar. *Berriman, mb4@sanger.ac.uk. *Nwaka, NwakaS@who.int. *Ralph, saralph@unimelb.edu.au. *Roos, droos@sas.upenn.edu. *Van Voorhis, wesley@u.washington.edu.

¹²Current address, University of Dundee, Dundee (UK)

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.cg>

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>

diseases are available, their use is often limited by factors including high cost, low efficacy, toxicity issues and the emergence of resistance¹.

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 encouraging¹. 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 datasets that are now available, filter out and prioritize a short list of candidate targets 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*, the malaria parasites *Plasmodium falciparum* and *P. vivax*; 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*, responsible for kala-azar and other forms of leishmaniasis, sleeping sickness, and Chagas disease, respectively (see Table 1 and Supplementary Methods).

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 datasets 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 datasets, 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, i.e., without optimization for specific targets. Notable exceptions include alpha-difluoromethylornithine (DFMO), which inhibits ornithine decarboxylase, for African sleeping sickness²; HIV protease inhibitors³; and zanamivir and oseltamivir, which inhibit the neuraminidase enzymes of influenza virus⁴. The ability to rapidly and effectively locate, capture, integrate, query and retrieve genomic-scale datasets 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.

Database content

The potential of a given pathogen gene product as a therapeutic target can be considered to depend 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 modulates that target into a drug. 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 and Supplementary Methods). 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 genome sequence has recently been published⁵, and at least a dozen worm genome projects are in progress. Comparison of sequences from multiple related *Plasmodium*, *Leishmania* and *Trypanosoma* species improve 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 produced sequence data to include new genomes as they become available, and to streamline data loading so that data 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 any of these species except *M. tuberculosis*^{6–9}. Thanks to our 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 ortholog identification methodology¹⁰, we have used OrthoMCL¹¹ to map all genes from targeted pathogens to orthologs in organisms for which additional useful information is available. High-quality whole-genome inactivation data are available for *S. cerevisiae* (in the form of genetic lesions^{12–14}) and *C. elegans* (in the form of RNA interference knockdowns¹⁵), and essentiality data on orthologues in bacteria¹⁶ 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 (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 any organism' is a useful criterion

for drug target prioritization as essential genes are more likely to be attractive candidates for drug development than those that are not essential in any known organism.

The TDR Targets database thus allows users to assess the role of the gene product in the life cycle 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 a homolog 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 of 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 druggability¹⁹ of a given target in a pathogen can be predicted on the basis of factors such as the physicochemical nature 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. Nevertheless, high-confidence structural models have been generated for many parasite proteins based on sequence similarity to orthologous proteins for which crystal structures are available²⁰. Gene pages in the TDR Targets database link to these models at ModBase. Potential inhibitors may also be inferred through orthology to proteins for which crystal structures or models of inhibitor-ligand interactions are available.

The database predicts the druggability of each parasite protein using various methods (see Supplementary Methods). Precedence for the druggability of targets was derived from Pfizer's comprehensive survey of known biological targets of drugs, leads and chemical tools²¹ and Biofocus's StARlite database (which is scheduled to become publicly available through the European Bioinformatics Institute (EBI)) of the medicinal chemistry literature. The sequences of approximately 1,400 proteins with known drug-like ligands were then mapped onto the 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 that was trained on a set of known drug targets¹². The druggability of each of the protein models generated for each parasite in the database was also assessed using a structure-based binding site algorithm¹³. A normalised, 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 homolog. A Compound Desirability Index (also ranging between 0 and 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 Methods). 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 targeted by specific compounds (see Supplementary Methods).

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 possible. In cases for which an 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 assay library. A total of 1,707 genes from the target organisms currently have such links to published assays. Enzyme assay development 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.

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. These requirements will differ among individuals and organizations, depending on experience with particular target classes and assay systems, the availability and diversity of compound libraries, strategies for hit identification, etc. Furthermore, different organisms will likely 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 parasites²² 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 variety of queries can be articulated, while prompting users to define parameters for potentially important criteria they may not have considered previously. Step-by-step examples of such searches are shown in Figures 1 and S2; all search results shown are based on data available as of July 2008.

Compilation of diverse data types 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. Intersection queries in TDRtargets.org can be conducted by choosing multiple features on the primary search page (see Figure 1 for an example), or can be generated by combining multiple queries 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 intersection queries can be powerful 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. Our 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 they consider 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 of all three criteria and head the list. The example in Box 2 and Supplemental Figure S2 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 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 upon previously published prioritization analyses. Published lists of promising targets for *M. tuberculosis*²³ and *Brugia* species²⁴ have been imported into the TDR Targets website and then posted for general use (see www.tdrtargets.org/published/browse/226 and www.tdrtargets.org/published/browse/230).

Challenges and future perspectives

The database was launched in March 2007 and as of July 2008 has attracted over 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 a variety of ways. For example, we have generated a manually curated list of enzymes (accessible via the Classification category of 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 TDR Targets 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 (Berriman *et al.*, unpublished), and the genomes of *Onchocerca volvulus* and several related nematode genomes (for example, *S. 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 the diagnosis of infection by TDR tropical organisms, we have recently added a 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). 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 the 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 per 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 networks²⁵ for both *in silico* and experimental analyses.

A major challenge in the coming years is to secure resources for the continued curation and further optimization of the database. Such optimization is likely to include the adaptation of the database to accommodate improved models for predicting druggability, gene-disruption datasets, the challenges of prioritizing drug targets in multicellular versus unicellular target organisms, and to the rapidly growing availability of medium and high-throughput compound screening 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 when carefully organized, integrated and annotated, and queryable in user-friendly environments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to acknowledge all of the investigators who provided the data in the TDR Targets database including those who 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 Preeti Bais and Barry Coflan for work on the association of targets with compounds; Rick L. Stevens (Argonne National Laboratory) for providing data for gene essentiality in bacteria; Kshitiz Chaudhary and Tilde Carlow (New England BioLabs) for integrated *C. elegans* phenotype data; Jim Sacchetti (Texas A&M) for information on known *M. tuberculosis* drug targets; and Mark Schreiber (Novartis Institute for Tropical Diseases, Singapore) and James 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), 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/NIAID). 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.

References

1. Nwaka S, Hudson A. Innovative lead discovery strategies for tropical diseases. *Nat Rev Drug Discov.* 2006; 5:941–55. [PubMed: 17080030]
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.* 2007; 33:359–66. [PubMed: 17610127]
3. Mori M, et al. Contribution of structural biology to clinically validated target proteins. *Drug Discov Today.* 2008; 13:469–472. [PubMed: 18549971]
4. Varghese JN. Development of neuraminidase inhibitors as anti-influenza virus drugs. *Drug Dev Res.* 1999; 46:176–196.
5. Ghedin E, et al. Draft genome of the filarial nematode parasite *Brugia malayi*. *Science.* 2007; 317:1756–60. [PubMed: 17885136]
6. McAdam RA, et al. Characterization of a *Mycobacterium tuberculosis* H37Rv transposon library reveals insertions in 351 ORFs and mutants with altered virulence. *Microbiology.* 2002; 148:2975–86. [PubMed: 12368431]
7. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol.* 2003; 48:77–84. [PubMed: 12657046]
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 U S A.* 2003; 100:7213–8. [PubMed: 12775759]
9. McNeil LK, et al. The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation. *Nucleic Acids Res.* 2007; 35:D347–53. [PubMed: 17145713]
10. Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 2006; 34:D363–8. [PubMed: 16381887]
11. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 2003; 13:2178–89. [PubMed: 12952885]
12. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? *Nat Rev Drug Discov.* 2006; 5:993–6. [PubMed: 17139284]
13. Al-Lazikani, B., et al. Bioinformatics - From Genomes to Therapies. Volume 3: The Holy Grail: Molecular Function. Lengauer, T., editor. Wiley-VCH; Weinheim: 2007. p. 1315-1334.
14. Winzler EA, et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science.* 1999; 285:901–6. [PubMed: 10436161]
15. Kamath RS, et al. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature.* 2003; 421:231–7. [PubMed: 12529635]
16. Gerdes SY, et al. Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J Bacteriol.* 2003; 185:5673–84. [PubMed: 13129938]
17. Titus RG, Gueiros-Filho FJ, de Freitas LA, Beverley SM. Development of a safe live *Leishmania* vaccine line by gene replacement. *Proc Natl Acad Sci U S A.* 1995; 92:10267–71. [PubMed: 7479765]
18. Chaudhary K, et al. Purine salvage pathways in the apicomplexan parasite *Toxoplasma gondii*. *J Biol Chem.* 2004; 279:31221–7. [PubMed: 15140885]
19. Hopkins AL, Groom CR. The druggable genome. *Nat Rev Drug Discov.* 2002; 1:727–30. [PubMed: 12209152]
20. Pieper U, et al. MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* 2004; 32:D217–22. [PubMed: 14681398]
21. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL. Global mapping of pharmacological space. *Nat Biotechnol.* 2006; 24:805–15. [PubMed: 16841068]
22. Meissner M, Breinich MS, Gilson PR, Crabb BS. Molecular genetic tools in *Toxoplasma* and *Plasmodium*: achievements and future needs. *Curr Opin Microbiol.* 2007; 10:349–56. [PubMed: 17826309]

23. Hasan S, Daugelat S, Rao PS, Schreiber M. Prioritizing genomic drug targets in pathogens: application to *Mycobacterium tuberculosis*. *PLoS Comput Biol*. 2006; 2:e61. [PubMed: 16789813]
24. Kumar S, et al. Mining predicted essential genes of *brugia malayi* for nematode drug targets. *PLoS ONE*. 2007; 2:e1189. [PubMed: 18000556]
25. Hopkins AL, Witty MJ, Nwaka S. Mission possible. *Nature*. 2007; 449:166–9. [PubMed: 17851512]
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*. 2003; 278:17075–83. [PubMed: 12618430]

Box 1 Definition of important criteria used in the database

Orthologues and Paralogues

Refer to homologous genes that arise either by speciation from a common ancestor (in case of orthologues) or by gene duplication (in case of paralogues). True orthologues tend to be functionally conserved whereas paralogues, arising from ancient duplication, can be functionally divergent. Ortholog groups are built by clustering orthologous proteins from multiple taxa. Such groupings provide the framework for comparing genes across multiple species and hence provide information for the functional annotation of genes. For example, in 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 the OrthoMCL database website.

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 Desirability

Provides a measure of the chemical quality of inhibitors that target a gene (or most similar gene) of interest. Compounds having 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 the 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 Supplementary methods.

Box 2 A weighted union of multiple queries to generate a ranked list of putative drug targets in *T. brucei*

Search criteria and weights

<i>Criterion</i>	<i>Weight</i>
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

Sample of ranked list of genes (top 24 genes, ranked by weighting above, of 10965 genes)

<i>Gene product</i>	<i>Gene name</i>	<i>Weight</i>
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	Tb1 1.02.0330	365
ornithine decarboxylase	Tb1 1.01.5300	365
cysteine peptidase C (CPC), 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	Tb1 1.02.0640	340
protein kinase, putative	Tb927.8.5950	340
receptor-type adenylate cyclase GRESAG 4	Tb1 1.03.0970	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 this example, the highest scoring target is farnesyl pyrophosphate synthase, a protein that additional experimental work suggests is a promising drug target²⁶. 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 2 structures for this enzyme in the Protein Data Bank as well as a ModBase model.
- The database has curated 12 bibliographic 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 Figure S2 for a pictorial summary of this multiple-query search.

TDR Targets Database www.TDRtargets.org

Identification and ranking of targets against neglected tropical diseases

home | search | history | posted target lists

Search for genes/targets

Use this form to search for proteins/targets using the following criteria.

1. Select a pathogen species of interest

☒ *Mycobacterium tuberculosis* ☐ *Leishmania major* ☐ *Trypanosoma brucei* ☐ *Trypanosoma cruzi* ☐ *Trypanosoma evansi* ☐ *Trypanosoma rangeli* ☐ *Trypanosoma vivax* ☐ *Trypanosoma evansi* ☐ *Trypanosoma rangeli* ☐ *Trypanosoma vivax*

2. Filter targets based on

Name/Annotation ☐ Search for targets using keywords (names, identifiers, functions, etc.)

Features ☐ Search based on key properties/features (molecular wt, isoelectric pt, length, etc.)

Structures ☐ Search for targets with associated 3D structures and/or models

Phylogenetic distribution ☐ Search for targets based on their phylogenetic pattern (species distribution)

Essentiality ☐ Search for targets that are essential for pathogen viability

Assayability ☐ Search for targets based on the availability of published assays

Druggability ☐ Search based on precedence and promise as a small molecule target

Curated targets ☐ Search for validated targets highlighted by manual curation

Bibliographic references ☐ Search for targets associated with publications indexed in PubMed

25 records found for *Plasmodium falciparum* source: PlasmoDB

Name	Ortholog group	Species
PFB0420w	OrthoMCL1_4173	2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
PFB0505c	OrthoMCL1_4172	Beta-ketacyl carrier protein synthase III
PFB0906c	OrthoMCL1_2654	Pandectin synthase
PFB0908w	OrthoMCL1_4001	Holo-lactyl carrier protein synthase
PFB1050c	OrthoMCL1_4390	4-diphosphocytidyl-2C-methyl-D-erythritol kinase (CMK)
PFB1070c	OrthoMCL1_1171	Uridylate kinase
PFB1030c	OrthoMCL1_2632	Phosphomethylpyrimidine kinase
PFB1070c	OrthoMCL1_4181	Enoyl acyl carrier reductase
PFB1050c	OrthoMCL1_2662	Choline synthase
PFB1070c	OrthoMCL1_3626	Cysteine desulfurase
PFB1070c	OrthoMCL1_3816	Pyruvate chain release factor
PFB1070c	OrthoMCL1_2153	Chlorophyll synthase
PFB1070c	OrthoMCL1_3606	Para-aminobenzoic acid synthase
PFB1070c	OrthoMCL1_4397	GcpB protein
PFB1070c	OrthoMCL1_2661	DNA Gyrase A-subunit
PFB1070c	OrthoMCL1_664	CoX-like hydrolase, hsd-superfamily, subfamily 1b
PFB1070c	OrthoMCL1_2521	DNA Gyrase B-subunit
PFB1070c	OrthoMCL1_4178	Beta-hydroxyacyl-CoA thiolhydratase
PFB1070c	OrthoMCL1_6746	Nucleoside transferase
PFB1070c	OrthoMCL1_3228	Apurinic/apyrimidinic endonuclease Apg1
PFB1070c	OrthoMCL1_6668	Phosphoenolpyruvate carboxylase
PFB1070c	OrthoMCL1_4442	ATP-dependent transporter
PFB1070c	OrthoMCL1_517	NAD(PH)-dependent glutamate synthase
PFB1070c	OrthoMCL1_6141	Serine/threonine phosphatase
PFB1070c	OrthoMCL1_3655	1-deoxy-D-xylulose 5-phosphate reductoisomerase

Name / annotation

Search for targets using keywords (names, identifiers, functions, text terms, etc.)

Identifier / Accession

EC number

Gene Ontology

Functional class

Plant/interpro domain

Structures

Search for targets with associated 3D structures and/or models

Retrieve targets with three dimensional data from:

PDB ☐

and/or

Modbase ☐

Phylogenetic distribution

Search for targets based on their phylogenetic pattern (species distribution)

Restricted to targets with orthologs present / absent in:

☐ *Escherichia coli* ☐ *Mycobacterium tuberculosis*

☐ *Mycobacterium leprae* ☐ *Plasmodium vivax*

☐ *Toxoplasma gondii* ☐ *Trypanosoma brucei*

☐ *Trypanosoma cruzi* ☐ *Trypanosoma rangeli*

☐ *Saccharomyces cerevisiae* ☐ *Drosophila melanogaster*

☐ *Brugia malayi* ☐ *Caenorhabditis elegans*

☐ *Mus musculus* ☐ *Homo sapiens*

Essentiality

Search for targets that are essential for viability, based on genomic-scale knock-out or knock-down studies (or orthologs of essential genes in other species)

☐ any evidence of essentiality in any species

or select the species and phenotype from options shown below (choosing more than one returns UNION of all options selected):

M. tuberculosis ☐

E. coli ☐

S. cerevisiae ☐

C. elegans ☐

My Queries:

☐ 1: *P. falciparum* w/EC or enzyme annotation, 862 records.

☐ 2: *P. falciparum* X-ray structures or high quality models, 4829 records.

☐ 3: *P. falciparum* ortholog of essential genes, 1162 records.

Combine or act on selected queries:

☐ Union: genes that appear in any of the selected queries.

☐ Intersection: genes that appear in all selected queries.

☐ Subtraction: subtract selected queries from first query.

☐ Delete: delete selected queries.

☐ Rename: rename selected queries.

☐ Save: save your selected queries as a set named

☐ Change species: view selected queries for

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 left panel). Next, one can query on function-based information. For example, selecting Enzyme under the Functional classification 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 the 4,829 genes for which there are crystal structures and/or high-quality structural models. Selecting within the “Phylogenetic distribution” panel (middle right) for orthologs present in *Plasmodium vivax* but not in humans or mice leads to 2,836 genes conserved in *Plasmodium* 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 orthologs 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 ortholog being essential in some model organism). There are many promising drug targets in *P. falciparum* among this list (circled in red), supporting this approach as one means to prioritize promising targets.

Figure 2. Ranking of *Mycobacterium tuberculosis* targets using the TDR Targets database

This Venn diagram demonstrates the gene-ranking functionality of TDR Targets database. This example shows the weighted union of queries to yield a list of ranked genes from *M. tuberculosis*. The selected criteria, the arbitrary weights assigned to each criterion 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 8 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 20) – 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 at the top. Note that many of these highly ranked genes code for cytochrome P450 enzymes, which do of course 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 ortholog groups. This suggests potential target selectivity.

Table 1

Selected lines of evidence accessible via the TDR Targets database

	Mycobacterium tuberculosis	Plasmodium falciparum	Leishmania major	Trypanosoma brucei	Trypanosoma cruzi
Enzymes	1790	862	2100	1918	5474
Assayability	192	205	255	283	252
Structures crystals models	229 2756	102 4829	37 3973	80 7119	49 7103
Phylogeny					
not in mammals	3294	3850	6606	8494	20730
Essentiality					
any evidence	1054	1162	1482	1375	2313
Phenotype data					
curated	93	107	171	301	–
Druggability					
Dindex >0.6	16	127	230	205	325
Desirability >0.6	31	23	37	33	54
Compound Links					
Curated	56	70	129	20	–
Drug Bank	–	115	260	287	477
PubMed (CAS+EC)	590	351	352	533	435

Table 2

Ranked list of *Mycobacterium tuberculosis* targets: top 50 (of 1,563 total 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, alpha subunit (DnaE1)	160
Rv1284	Conserved hypothetical protein	240	Rv1601	Imidazole glycerol-phosphate dehydratase (HisB)	160
Rv1730c	Penicillin-binding protein	170	Rv2268c	Cytochrome P450 Cyp128	160
Rv0533c	Beta-ketoacyl-ACP synthase III (FabH)	160	Rv2157c	UDP-N-acetyl-muramoyl-alanyl-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-acetyl-muramoyl-pentapeptidyltransferase (MurX)	160	Rv0886	NADPH:adrenodoxin oxidoreductase (FprB)	150
Rv2952	Methyltransferase	160	Rv1348	ABC Transporter	150
Rv3711c	DNA Polymerase III, epsilon subunit (DnaQ)	160	Rv1714	Oxidoreductase	150
Rv3804c	Secreted antigen 85 complex A (FbpA)	160	Rv2122c	Phosphoribosyl-AMP pyrophosphatase HisE	150
Rv2068c	Class A beta-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 End	150
Rv0778	Cytochrome P450 Cyp126	160	Rv1463	ABC Transporter	150
Rv2540c	Chorismate synthase (AroF)	160	Rv2220	Glutamine synthetase GlnA1	150
Rv3803c	Secr. 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