Supplementary box S1 | Methods for TDRtargets.org

Database infrastructure. TDRtargets.org is driven by a relational database and an associated web application, both designed in a modular fashion to facilitate incorporation of new datasets or implementation of new queries. The database is managed by MySQL and consists of ~30 tables used to store and relate information. The web application is written in Perl, taking advantage of various frameworks (Catalyst, DBIx::Class, Template Toolkit). In sum, the web application consists of code for interfacing with the database, controller code managing interaction with the user and implementing functionalities (such as query history and weighting) and templates for web page production. The system runs on dedicated FreeBSD servers.

Genome datasets. Genome sequence data, gene predictions and the conceptually translated proteome were imported from public databases. *Plasmodium* information was obtained from Plasmodb.org. Information for *Trypanosoma brucei*, *T. cruzi* and *Leishmania major* was imported from GeneDB.org. *Mycobacterium tuberculosis* genome information was imported from TubercuList. Identification of protein domains (Interpro), assignment of Gene Ontology (GO) terms and enzyme commission (EC) numbers were obtained from these data providers or generated by using InterProScan. Functional class assignment was based on high level GO terms and/or metabolic pathways from the GO 'molecular function' ontology.

Functional genomics data. Expression data (abundance percentiles) were obtained for persisting, dormant-stage *Mycobacterium tuberculosis* and various stages of *Plasmodium falciparum* (from Plasmodb.org). Incorporation of additional functional genomics data, for other stages and species, is in progress.

Functional Category. Genes were classified into enzyme, transporter and receptor categories as follows:

**Enzyme:** genes that have one or more of the following features: 1) an EC number; 2) a GO term for catalytic activity (GO:0003824), or one of its more specific subterms e.g. kinase activity (GO:0016301); 3) annotation with an enzyme name, e.g. dehydrogenase, calpain etc; 4) orthology to known enzymes from other organisms (e.g. *Saccharomyces cerevisiae*).

**Transporter:** genes that have one or more of the following features: 1) a GO term for transporter activity (GO:0005215) or one of its more specific subterms, e.g. chloride channel activity (GO:0005254) (excluded any genes associated with non-transmembrane transport, e.g. carrier proteins or proteins involved in vesicle transport); 2) annotation with a transporter name eg pteridine transporter.

**Receptor:** genes that have one or more of the following features: 1) a GO term for receptor activity (GO:0004872) or one of its more specific subterms, e.g. peptide receptor activity (GO:0001653); 2) annotation with "receptor" in their name.

Metabolic pathway mapping. Under this category, pathway information available for individual genes is captured. Data was obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) in which genes are mapped to pathways using EC and ortholog information. KEGG contains both very specific pathways (e.g. purine metabolism) as well as broader categories that encompass the specific pathways (e.g. purine metabolism).
nucleotide metabolism); the former are referred to in TDR Targets as the detailed pathways whereas the latter are referred to as the high-level pathways.

**Assayability data.** Information about whether or not an enzyme can be assayed was obtained from the Sigma Aldrich assay library. Assayability information was mapped based on the EC number annotation of target genes.

**Structural data.** Structural data for specific taxa of interest (or closely related species) were obtained from the Protein Data Bank (PDB.org); for example, the structure of topoisomerase I from *Leishmania donovani* (PDB #2B9S) is linked to *L. major* topoisomerase I (GeneDB #LmjF04.0060) in the TDRtargets database. Structural predictions produced by homology-based modeling were obtained from ModBase.compbio.ucsf.edu.

**Phylogenetic distribution.** In order to facilitate queries based on species distribution, phylogenetic distribution patterns were defined by ortholog groups obtained from OrthoMCL, which identifies paralogs and orthologs using a reciprocal best hit algorithm and Markov clustering. These groups were also used to facilitate orthology-based inference (see below).

**Antigenicity.** Antigenic peptides were predicted for proteins in the database using the method of Kolaskar and Tongaonkar as implemented in the EMBOSS package. Each predicted epitope is scored based on the physicochemical properties its amino acid residues and a cumulative antigenicity index, allowing users to formulate queries based on the predicted antigenicity relative to other proteins encoded by the pathogen genome.

**Essentiality data.** Information on gene essentiality derived from large-scale functional studies (mostly in non-target species) was obtained from published studies and online data resources, including the *Saccharomyces* Genome Database YeastGenome.org, the Profiling *E. coli* Chromosome database shigen.nig.ac.jp/ecoli/pec/, the Keio Collection Ecoli.naist.jp, the National Microbial Pathogen Data Resource NMPDR.org, and the *C. elegans* genome database WormBase.org. Additional essentiality information on *C. elegans* was kindly provided by Drs. T. Carlow and K. Chaudhary of New England Biolabs. Essentiality data from these various species was linked to parasite genes by orthology.

**Druggability.** The druggability of each parasite protein was assessed by using several different methods, each with varying degrees of confidence. Precedence for the druggability of targets was derived from Pfizer's comprehensive survey of the known biological targets of drugs, leads and chemical tools as well as the analysis of two highly curated Biofocus DPI chemogenomics databases: DrugStore and StARLiTe. DrugStore is a database of FDA approved drugs, their efficacy targets and therapeutic indications, and contains about 380 molecular targets and ~1500 therapeutic entities. The efficacy targets were expert curated from published literature and FDA drug literature. StARLiTe is a database of abstracted medicinal chemistry literature; it contains >350,000 compounds that have been screened in biological systems, and >3,500 protein molecular targets with screening data.
A number of measures of druggability were calculated, both homology-based and homology-independent. Pathogen genome sequences were mapped to chemically characterized proteins using high-confidence orthology mapping as well as simple BLAST sequence similarity searches. Three sets of chemically characterized proteins were compiled from DrugStore and StARLITE to reflect three decreasing degrees of useful preceeded chemistry: i) the list of approved drug targets (all drug targets from DrugStore), ii) the list of ~1,400 literature targets that show acceptable affinity (≤ 2 µM) when screened with drug-like compounds (as defined by Lipinski\cite{Lipinski2001}), and finally iii) the list of literature targets that show acceptable affinity (≤ 2 µM) when screened with any compound regardless of its quality.

The non-homology-restricted feature-based druggability was also calculated using a sequence feature-based Bayesian algorithm trained on a set of known drug targets\cite{Zhou2004}. Additionally, the structure-based druggability of each of the protein models generated for each of the parasite proteomes in the database were assessed using a structure-based algorithm\cite{Briggs2005}.

A normalised, weighted sum based on the accumulation of prediction for the different methods results in a composite 'Druggability Confidence' index (value range of 0 – 1 with 1 as the ideal value) for each parasite protein. The index also reflects the degree of similarity between the pathogen target and the known druggable homolog. Based on this analysis, a large number of proteins from the TDR priority organisms could be linked to known targets with at least one small molecule compound with a binding affinity of less than 2 µM.

**Compound Desirability Index.** A compound desirability index was assigned to each target gene based on the chemical quality of known inhibitors of the most similar target in other species, incorporating Harrington’s desirability index\cite{Harrington1995,Harrington1996} (addressing molecular properties and oral distribution of small molecule drugs), with penalties for acidity, promiscuity, and structural alerts (reactive and toxic groups). By linking predicted druggable targets to orthologs and homologs with known chemical matter\cite{Kurth2005}, it is also possible to provide a metric summarizing the average chemical quality of ligands as a start for drug development. The median of the sum of desirabilities for all ligands for a target provides a 'Target Compound Desirability' value (value ranges from 0 – 1), with the ideal value being 1.

**Chemical compounds.** Although this database focuses on the identification and prioritization of drug targets from a gene/protein-centric approach, links to chemical compounds are also included, providing the ability to search for genes targeted by specific compounds. Although proprietary data derived from the compound desirability analysis (above) cannot be disclosed, information on potentially relevant compound-target associations was collected by manual curation of the literature (see below), transitive assignment based on orthology or sequence similarity to proteins with compound links in DrugBank (redpoll.pharmacy.ualberta.ca/drugbank), and transitive assignment based on the co-occurrence of EC numbers and CAS registry numbers in published PubMed abstracts. A total of 3764 gene products were associated with compounds (315 via curated gene linkages, 1139 via DrugBank, and 2310 via PubMed).

**Curated annotation.** Phenotypic data for *L. major* and other *Leishmania* spp., *T. brucei*, *P. falciparum*, and *M. tuberculosis* was obtained from literature and web searches and captured using a combination of controlled vocabularies including GO
and the evidence ontology (GeneOntology.org). Information extracted from bibliographic sources was used to describe what was altered (growth, morphology, catalytic activity), the nature of the alteration (normal, abnormal, reduced, slow), timing (e.g. life cycle stage), the method or evidence employed (e.g. loss-of-function mutant, specific protein inhibition), and the bibliographic source. Information on proposed drug targets for *T. brucei* was also contributed by members of the scientific community through a survey conducted in 2007 (www.tdrtargets.org/survey).

References
