Supporting Information

Plouffe *et al.* 10.1073/pnas.0802982105

SI Materials and Methods

Screening Media and Complete Media. Screening medium contained no human serum. Human serum was considered a potential source of variability between screens, and using only albumax in the screening media produced more consistent results, and cost less. Screening medium consisted of RPMI (without phenol red, with L-glutamine), 4.16 mg/ml albumax II, 0.013 mg/ml hypoxanthine, 1.73 mg/ml glucose, 0.18% $NaHCO_3$, 0.031M Hepes, 2.60 mM NaOH, and 0.043 mg/ml gentamicin. Complete medium consisted of RPMI (without phenol red, with L-glutamine), 4.3% heat-inactivated O+ human serum, 2.08 mg/ml albumax, 0.013 mg/ml hypoxanthine, 1.17 mg/ml glucose, 0.18% $NaHCO_3$, 0.031 M Hepes, 2.60 mM NaOH, and 0.043 mg/ml gentamicin.

Blood Preparation. Before each screen, three units of O+ whole blood from donors between the ages of 18 and 39 were obtained from Interstate Blood Bank. The blood was tested for HIV, HCV, and other pathogens. Leukocytes were removed from the blood by using a Sepacell filtration device (Baxter International). The blood was centrifuged at 2,000 rpm for 5 min using an Eppendorf centrifuge 5810R with an A-4-81 rotor and washed with an equal volume of complete medium. This was repeated three times, and the blood was stored at a 50% hematocrit at 4°C.

Blood Testing. The quality of the blood was tested before any screening to ascertain the parasitic infection rate and potential signal range. Samples from each unit of blood were centrifuged down at 2,000 rpm for 5 min using an Eppendorf centrifuge 5810R with an A-4-81 rotor, and the complete medium was aspirated off. In screening medium, for each blood sample, a suspension was made at a 0.3% parasitemia and 2.5% hematocrit with the 3D7 strain of *P. falciparum*. In a 384-well, black, clear-bottom plate (Greiner), 50 μl of from each suspension was dispensed into two columns. From each set, the first column was untreated and would serve as a baseline value, and the second column was treated mefloquine at a final concentration of 10 μM and would completely inhibit parasite proliferation and serve as a background value. The plate was incubated at 37°C for 72 h in an environment of 93% nitrogen, 4% carbon dioxide, and 3% oxygen. Ten microliters of detection reagent consisting of 10× SYBR Green I (Invitrogen; supplied in 10,000× concentration) in lysis buffer (20 mM Tris·HCl, 5 mM EDTA, 0.16% Saponin wt/vol, 1.6% Triton X vol/vol) was added to the wells. The plate was sealed with foil and left at room temperature for 24 h. The plate was centrifuged at 700 rpm for 30 sec, using an Eppendorf centrifuge 5810R with an A-4-81 rotor and the foil was removed. The fluorescence intensity from the plate was read by using an Analyst GT multimode reader (Molecular Devices); the reader required a 505 dichroic mirror with 485-nm excitation and 530-nm emission settings, and the plate was read from the bottom. For each blood sample, the average baseline value was divided by the average background average. Typically, the baseline value was a 6- to 10-fold higher than the background. The blood that produced that highest signal range was used for the screen. No blood with a <6-fold signal range and older than 2 weeks after the draw date was used for screening.

Genomics Institute of the Novartis Research Foundation (GNF) Chemical Library and Diversity Analysis. The GNF chemical library consists of >1.7 million compounds. At the time of the screening, most compounds were guaranteed to be at least 85% pure. They have drug-like characteristics with a high observance to Lipinski's rule of five.

Each compound structure was represented by a binary 512-bit ChemAxon 2D chemical fingerprint, and structural similarity was measured by Tanimoto metric as described in ref. 1. To evaluate the statistical significance of chemical similarities, a million random pairs of structures were sampled from the compound library and their Tanimoto scores calculated. Structures in the GNF screening collection have an average similarity score of 0.37 ± 0.10, nearly identical to the PubChem library (0.39 ± 0.11). For a given similarity score, a *P* value was assigned as the proportion of the above samples with equal or better scores.

Antimalarial Proliferation Screen in a 1,536-Well Plate Format. There were two aspects of the antimalarial proliferation screen that were unconventional, the first being that the parasites required a low-oxygen atmosphere to grow, requiring the production of special units, and the second being that reading the assay plate from the bottom yielded a 3- to 4-fold higher signal range than reading the assay plate from the top and that the on-line system was not equipped for bottom reads. Because of the special atmospheric requirements for *Plasmodium* proliferation, and the fact that bottom reads produce a several-fold higher fluorescence signal window, the screen was divided into four parts, and a hybrid screening approach was taken in that the dispensing steps were performed on-line by using GNF's on-line integrated screening system (GNF Systems), and the incubation and assay plate reads were performed on off-line workstations.

The 3D7 strain of *P. falciparum* was cultured in complete medium (2) until the parasitemia reached 3–8%. Parasitemia was determined by checking at least 500 red blood cells from a Giemsa-stained blood smear. The 3D7 cultures along with tested O+ red blood cells were centrifuged at room temperature at 2,000 rpm for 5 min using an Eppendorf centrifuge 5810R with an A-4-81 rotor. The medium was aspirated off. For the compound screening, a parasite dilution at a 0.3% parasitemia and 4.0% hematocrit was created with screening medium. The suspension was gassed with 93% nitrogen, 4% carbon dioxide, and 3% oxygen and placed at 37°C until needed.

Using GNF on-line screening equipment, 3 μl of screening medium was dispensed into 1,536-well, black, clear-bottom plates (Greiner). With a PinTool (GNF Systems), 10 nl of compounds was transferred into the assay plates along with control compounds. Next, 5 μl of the parasite suspension in screening medium (see above) was then dispensed into the assay plates such that the final parasitemia was 0.3% and the final hematocrit was 2.5%. Compounds were screened at a 1.25 μM concentration with a final DMSO concentration of 0.125%. Mefloquine at a final concentration of 12.5 μM and DMSO at a final concentration of 0.125% were used within the assay plates to serve as background and baseline controls, respectively. The assay plates were transferred to off-line incubators that contained airtight incubation units. The units were gassed daily with 93% nitrogen, 4% carbon dioxide, and 3% oxygen during the 72-h incubation at 37°C. Two microliters of detection reagent consisting of 10× SYBR Green I (Invitrogen; supplied in 10,000× concentration) in lysis buffer (20 mM Tris·HCl, 5 mM EDTA, 0.16% Saponin wt/vol, 1.6% Triton X vol/vol) was dispensed into the assay plates. For optimal staining, the assay plates were left at room temperature for 24 h in the dark. The assay plates were read off-line by using several Acquest GT

multimode readers (Molecular Devices); the readers required 505 dichroic mirrors with 485-nm excitation and 530-nm emission settings, and the plate reads were from the bottom.

Compound Activity Matrix. The GNF has routinely screened up to 1.7 million compounds in >200 biological assays over the years, which results in >130 million single-dose activity data points in its Lead Discovery Database. For this study, all HTS data for each of the 17,074 primary screening hits were retrieved from the database into a large compound-by-assay activity matrix. When a missing value was encountered, the measurement from other wells containing the same structure for the same assay, if available, was used instead. Assays covering <10% of the hits were then removed, and 131 assays remained. Compounds containing measurements in <50% of the remaining assays were removed, resulting in 8,457 compounds. For antagonist assays, activity data were already in the form of percent inhibition ranging between 0% and 100%. For agonist assays, reciprocal of percent induction was used to map the data into the same range, where 0% represents no induction and 50% for a 2-fold induction.

Structure Clustering of Hits. Compounds were first hierarchically clustered; subtrees above the cutoff were collapsed into individual structure classes that typically share a common scaffold. SAR principle is generally considered effective for compound structures sharing a Tanimoto similarity score of 0.85 or higher (3); therefore, this was used as a similarity cutoff in this study. Structure clustering was applied to both primary hits and validated hits independently. In the chemotype analysis (Fig. 2), a confirmed structure class was counted as one new chemical starting point only when no association to any known drug could be found for any of its class member.

Knowledge Annotation of Compounds. GNF has established an automated informatics pipeline that compiles existing knowledge for small molecules from various public databases such as NCBI PubChem and licensed databases such as the World Drug Index (4). The in-house compound annotation database currently consists of various annotations for ≈800,000 unique structures. This database enabled us to assign annotations to an HTS hit if it matches a database entry with similarity score of 0.95 or higher. Among all of the databases, NCBI PubChem directly links compound structures to the MeSH pharmacological action ontology, which essentially provides a set of controlled vocabularies to classify known drugs into different MOA categories.

Profile Analysis by Ontology-Based Pattern Identification Algorithm. Detailed description of the ontology-based pattern identification (OPI) algorithm can be found in previous bioinformatics and cheminformatics studies (5–7). Although compounds sharing similar activity profiles can be grouped together by many clustering techniques, the OPI algorithm is specifically designed to identify the optimal boundaries of local activity patterns shared by a group of compounds. Within the list those compounds belonging to a given targeted family are most statistically enriched. In particular, a metaprofile that best represents the known class members was constructed. The metaprofile could be taken either from that of a certain class member or from averaging profiles of several members. All compounds were then ranked according to their profile correlation with the metaprofile, under the assumption that compounds ranking toward the top were more likely to share the class label (scaffold or MOA). The subset of top-ranking compounds were predicted to either contain the scaffold, if they were structurally uncharacterized, or share the same MOA. The exact cutoff was determined by an iterative scoring procedure in such a way that the resultant compound list had the optimal enrichment of known class members.

The enrichment score is essentially a statistical *P* value, which represents the likelihood of obtaining such a tight cluster in both chemical/pharmacological and biological space by chance. When the target family is a group of compounds sharing a common scaffold (e.g., staurosporine-like structures), OPI identifies those chemical neighborhoods with conserved biological profiles (6), which are in general agreement with the SAR principle and serve as a desirable starting point for later medicinal chemistry exploration. In addition, the target family can also be a group of compounds sharing the same MOA (i.e., a MeSH group) but not necessarily close in chemical space. OPI analysis would then be able to identify unique biological profiles shared by compounds enriched in a particular MOA. When such an enrichment is statistically solid, the profile is treated as an HTS fingerprint for a certain MOA; i.e., it can be used to identify uncharacterized compounds of the same MOA with reasonable confidence. To ensure the statistical significance of the pattern identified by OPI, compounds in the activity matrix were randomly relabeled 100 times, and the same analyses were repeated. Only those *P* values reproducible by ≤5% simulations were considered truly significant by this study; i.e., $P_{\text{permute}} \leq 5\%$.

Docking Studies. *In silico* docking of the diaminopyrimidine derivative cluster into *P. falciparum* DHFR-TS (PfDHFR-TS) was performed with AutoDock 3.0.5 with AutoDockTools 1.4.5 (8). The protein target was generated from the crystal structure of wild-type PfDHFR-TS in complex with inhibitor WR99210 [Protein Data Bank (PDB) entry 1J3I]. Because of the bifunctional nature of the protein and the native dimeric conformation, only the coordinates for a single DHFR domain, constituted by chain A (residues 1–133), were relevant and retained as the docking target. All ligands, except for the NADPH cofactor, were removed, and the model was input to AutoDockTools (ADT) for preparation of the edited PDB file into pdbqt format. Similarly, the PDB files for each diaminopyrimidine derivative were converted to the appropriate pdbq format via ADT and submitted to AutoGrid with the default settings. A hybrid genetic algorithm with local search (GALS) was chosen as the docking algorithm in AutoDock, and an initial round of blind docking (0.503-Å grid spacing) was performed for each compound to the PfDHFR model to confirm the dihydrofolate (DHF) active site as the most favored binding site. A second round of docking was repeated by using a search space (0.375-Å grid spacing) confined to the DHF active site. The results of the AutoDock trials were analyzed by using ADT and then exported to PyMOL (9) for graphical rendering.

1. Yan SF, Asatryan H, Li J, Zhou Y (2005) Novel statistical approach for primary high-throughput screening hit selection. *J Chem Inf Model* 45:1784–1790.
2. Trager W, Jensen JB (1976) Human malaria parasites in continuous culture. *Science* 193:673–675.
3. Wilton D, Willett P, Lawson K, Mullier G (2003) Comparison of ranking methods for virtual screening in lead-discovery programs. *J Chem Inf Comput Sci* 43:469–474.
4. Zhou Y, *et al.* (2007) Large-scale annotation of small-molecule libraries using public databases. *J Chem Inf Model* 47:1386–1394.
5. Zhou Y, *et al.* (2004) In silico gene function prediction using ontology-based pattern identification. *Bioinformatics* 21:1237–1245.
6. Yan SF, *et al.* (2006) Learning from the data: mining of large high-throughput screening databases. *J Chem Inf Model* 46:2381–2395.
7. Zhou Y, *et al.* (2008) Evidence-based annotation of the malaria parasite's genome using comparative expression profiling. *PLoS ONE* 3:e1570.
8. Morris GM, *et al.* (1998) Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function. *J Comput Chem* 19:1639–1662.
9. DeLano WL (2002) The PyMOL Molecular Graphics System (DeLano Scientific, Palo Alto, CA).