ARTICLES

Thousands of chemical starting points for antimalarial lead identification

Francisco-Javier Gamo[1], Laura M. Sanz[1], Jaume Vidal[1], Cristina de Cozar[1], Emilio Alvarez[1], Jose-Luis Lavandera[1], Dana E. Vanderwall[2], Darren V. S. Green[3], Vinod Kumar[4], Samiul Hasan[4], James R. Brown[4], Catherine E. Peishoff[5], Lon R. Cardon[6] & Jose F. Garcia-Bustos[1]

Malaria is a devastating infection caused by protozoa of the genus *Plasmodium*. Drug resistance is widespread, no new chemical class of antimalarials has been introduced into clinical practice since 1996 and there is a recent rise of parasite strains with reduced sensitivity to the newest drugs. We screened nearly 2 million compounds in GlaxoSmithKline's chemical library for inhibitors of *P. falciparum*, of which 13,533 were confirmed to inhibit parasite growth by at least 80% at 2 μM concentration. More than 8,000 also showed potent activity against the multidrug resistant strain Dd2. Most (82%) compounds originate from internal company projects and are new to the malaria community. Analyses using historic assay data suggest several novel mechanisms of antimalarial action, such as inhibition of protein kinases and host–pathogen interaction related targets. Chemical structures and associated data are hereby made public to encourage additional drug lead identification efforts and further research into this disease.

With approximately 243 million cases and 863,000 attributed deaths reported globally in 2009 (ref. 1), malaria is one of the most severe infectious diseases, primarily affecting the world's most disadvantaged populations. Of the four typically recognized *Plasmodium* species causing disease in humans, *Plasmodium falciparum* causes most mortality, mainly in children below the age of 5, and *Plasmodium vivax* most morbidity, additionally representing a reservoir of latent infection that hampers current control and future elimination efforts[2]. No new class of antimalarials has been introduced into clinical practice since 1996 (ref. 3), owing to the intrinsic difficulties in discovering and developing new antimicrobials, as well as a relative lack of public and private resource commitment towards antimalarial research. Today, the last class of widely efficacious drugs, the artemisinins, is being compromised by the rise of *P. falciparum* strains with reduced clinical response to artemisinin-containing drug combinations[4–6]. The genomics revolution has not yet led to new antimalarial medicines and target-based lead discovery has produced disappointing results, generally for lack of whole-cell activity as documented for antibacterials[7]. To secure that property in all chemical starting points for new antimalarial leads, we have tested the approximately 2 million-compound library used for high throughput screening at GlaxoSmithKline (GSK) for inhibitors of *P. falciparum*'s intraerythrocytic cycle, the *Plasmodium* species causing the highest mortality and the parasite growth phase responsible for disease symptoms as well as being amenable to *in vitro* culture. Here we describe 13,533 compounds confirmed to inhibit parasite growth by more than 80% at 2 μM concentration. Only 15% displayed some cytotoxicity in that they inhibited proliferation of the HepG2 human hepatoma cell line by more than 50% at 10 μM. All of these proven plasmodial inhibitors, of which 82% were previously proprietary and thus unknown to the general research community, are hereby made public to accelerate the pace of drug development for malaria.

Tres Cantos antimalarial compound set (TCAMS)

The 1,986,056 compounds present in GSK's screening collection in January 2009 were tested for inhibition of *P. falciparum* 3D7 at 2 μM under *in vitro* conditions described in Methods. 19,451 primary hits inhibiting parasite growth by more than 80% were obtained. Fresh samples of these primary hits were tested in two independent experiments and compounds displaying 80% or higher inhibition of parasite growth in at least two of the three assay runs were considered confirmed hits. 13,533 compounds were identified using this protocol (confirmation rate $> 70\%$). We did not detect any compounds in this set as non-specific inhibitors of the biochemical readout system by testing directly for inhibition of lactate dehydrogenase (LDH) in *P. falciparum* extracts (Methods). Evidence of cytotoxicity against human hepatoma HepG2 cells (a widely used *in vitro* marker for liver toxicity[8]), or interference with the luciferase reporter system used in the cytotoxicity assay (Methods), was observed in just 1,982 of the compounds when tested at 10 μM. This relative lack of non-specific cell toxicity is probably due in part to the low (2 μM) primary screening concentration used[9]. Estimation of the concentrations producing 50% inhibition of *P. falciparum* growth (XC_{50}, see Methods) indicated that most compounds are sub-micromolar inhibitors. The full compound set (TCAMS) and data table (Supplementary Table 1 and available at http://www.ebi.ac.uk/chemblntd) contains 13,533 compound entries. We have detected 139 of these as variations in salt form or stereochemistry of 68 parent structures, which make good internal controls for the biological assay data. They appear as different compounds with the same structure. When the stereochemistry is resolved it shows in the SMILES structural code in Supplementary Table 1 and in the Chembl-NTD database (http://www.ebi.ac.uk/chemblntd).

Representatives from all but one class of clinically used antimalarials have been recovered in the screen, providing additional validation

[1]Tres Cantos Medicines Development Campus, GlaxoSmithKline, Severo Ochoa 2, 28760 Tres Cantos, Spain. [2]Computational and Structural Chemistry, GlaxoSmithKline, Five Moore Drive, Research Triangle Park, North Carolina 27709-3398, USA. [3]Computational and Structural Chemistry, GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Hertfordshire, Stevenage SG1 2NY, UK. [4]Computational Biology, Quantitative Sciences, GlaxoSmithKline, 1250 South Collegeville Road, Collegeville, Pennsylvania 19426, USA. [5]Computational and Structural Chemistry, GlaxoSmithKline, 1250 South Collegeville Road, Collegeville, Pennsylvania 19426, USA. [6]Quantitative Sciences, GlaxoSmithKline, 709 Swedeland Road, King of Prussia, Pennsylvania 19406, USA.

of the selected assay. Several analogues of 4-aminoquinolines (for example, chloroquine), 8-aminoquinolines (for example, primaquine), methanol-quinolines (for example, quinine), diaminopyrimidines (for example, pyrimethamine), diaminotriazines (for example, cycloguanil) and naphthoquinones (for example, atovaquone) are all present in the hit collection. They do not represent more than 10% of all the hits by our similarity criteria. Endoperoxides (for example, artemisinin) were not identified, as they are not present in the screening library.

In addition to screening the reference laboratory strain 3D7, TCAMS was subsequently screened against the multidrug resistant Dd2 strain at the same concentration used in the primary screen (2 μM). This strain is insensitive to several quinolines and antifolates. Results showed approximately 8,000 compounds (~60%) inhibited Dd2 growth by more than 50% (Supplementary Table 1). Compounds potent against 3D7 (XC_{50} lower than 200 nM), but less active against Dd2 (growth inhibition less than 50% at 2 μM), are predominantly quinolines or structures related to antifolates, as expected. But Dd2 has also been shown to be less sensitive to unrelated chemotypes[10], probably due to amplification and mutation of the efflux pump gene *pfmdr1* (ref. 11). Specific characterization of individual hits is required before more detailed conclusions can be drawn, but it is encouraging that more than half of our hits retain potent activity against a multidrug resistant strain.

Structural characterization of hits

We used two methods to structurally characterize TCAMS and estimate the number of different chemotypes present. The hits can be described by 416 molecular frameworks[12] or by fingerprint clusters. Here, we use the Daylight fingerprint methods with a Tanimoto similarity index of 0.85 (ref. 13) which yields 857 clusters and 1,978 'singletons' (three or fewer similar compounds) (Supplementary Table 1). These fingerprint cluster annotations can be used to separate distinct classes from within the broader molecular framework categories. This is because the molecular frameworks describe the core template of compounds well (and therefore provide a robust, consistent categorization of compounds) whereas the fingerprint methods will often capture subtle substituent patterns but miss the commonalities in the core; together the two methods allow an ordered navigation through the chemical structure space represented by TCAMS and exemplified graphically in Fig. 1. In the absence of specific target information, such chemical structure clustering provides a framework to enable a systematic exploration of the mode of action for the compounds. The assumption being that compounds in the same cluster share mode of action (similarity principle[14]). This may not always be true but it is the best starting hypothesis in the absence of additional information. Therefore, taking just a few exemplars from each cluster can reduce the work required to identify

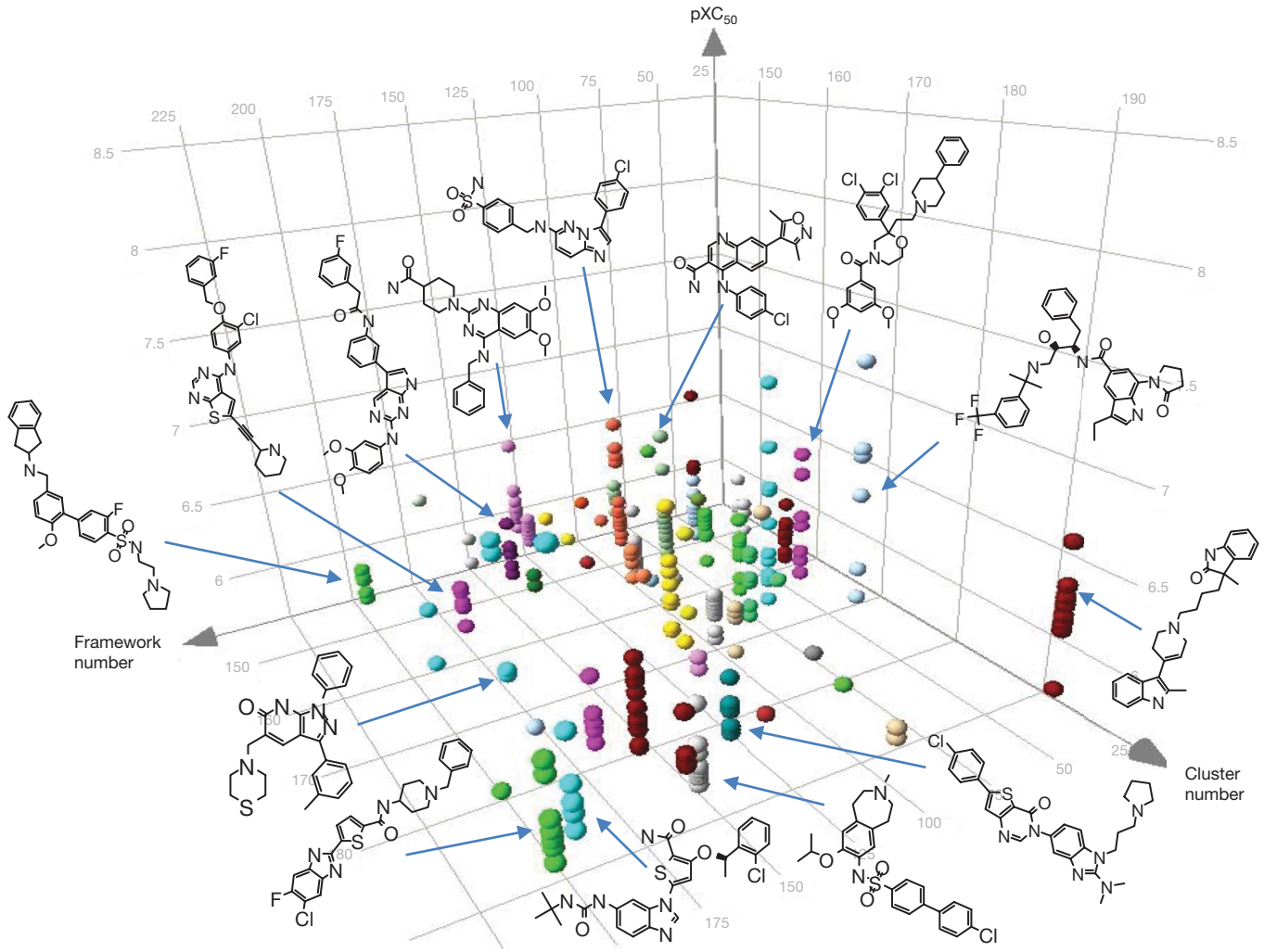


Figure 1 | Three-dimensional plot of some of the novel chemical diversity present in TCAMS. Compounds are represented by coloured spheres plotted against their assigned molecular framework number, chemical fingerprint cluster number and estimated antiplasmodial potency ($pXC_{50} = -\log XC_{50}$, where XC_{50} is in molar units and pXC_{50} is dimensionless; $XC_{50} = 1$ μM corresponds to $pXC_{50} = 6$). Inserted structures are examples of drug-like molecules not previously described to possess antiplasmodial activity.

specific targets for a set of molecules. Fingerprinting methods also enable searching outside of TCAMS for the activity of related molecules that can inform the biological activity of a given chemical cluster. Molecular framework and fingerprint cluster numbers are indicated for all compounds in Supplementary Table 1.

From a physicochemical point of view, our screen selected for compounds having a larger molecular mass and a higher hydrophobicity index than the average for the source compound collection (446 versus 385 Da; 5 versus 3.3 clogP, respectively; Fig. 2a, b). The meaning of this observation is unclear, but it may have to do with the physicochemical requirements needed to reach intracellular targets. The same reasons may also explain the bias towards compounds made for internal projects (82% of the hits) versus those commercially sourced (50% of the screening collection), as the latter tend to be smaller and more hydrophilic (commercial suppliers are listed in Supplementary Table 1). Database searches show that most internal compounds have not been published previously (see Supplementary Table 2 for exceptions), thus potentially represent highly novel chemical diversity made available to the malaria community. Some automatic methods may flag compounds in TCAMS with structural features a priori undesirable in drug leads[15], but because structural alerts are not always borne out and TCAMS compounds are not overtly cytotoxic we have retained them, because they can at least be useful for target identification and validation.

Modes of action hypotheses

We have used historical GSK data to develop testable hypotheses for the antimalarial mode of action of hits in an attempt to facilitate target identification efforts in the broader malaria community. We found 4,205 compounds with unambiguous annotations regarding their activity in biochemical assays against human (3,435 compounds) or microbial targets (770 compounds), some of which have been reported in the open literature (Supplementary Table 2). Inhibitors with highly promiscuous and non-specific profiles were identified by calculating an 'inhibition frequency index' (IFI) (see Methods) that is reported for every structure (Supplementary Table 1). These compounds were not used for analysis but remain in the data set. We also used conservative thresholds to consider a compound a true effector of a specific human target (that is, sub-micromolar 50% inhibitory concentration (IC_{50}), see Methods). Overall, activities at 413 targets (some assayed in both antagonist and agonist modes) were analysed among the antimalarial compounds. Figure 2c shows the distribution of known human target class activity among the hits. Approximately 70% are G-protein coupled receptors (GPCRs) or protein kinases. This reflects partly the relative abundance of different ligand classes in GSK's screening collection, and it would bias any straightforward prediction of the antimalarial mode of action. We therefore wanted to distinguish targets preferentially inhibited by TCAMS compounds from those equally prevalent among the compounds not present in TCAMS, under the hypothesis that the former were more likely to be human orthologues of the lethal target in *Plasmodium*. We calculated an 'enrichment' factor as simply the ratio of target actives among hits, over a background defined as all target actives among all screened compounds with data for that target (see Methods). An enrichment factor of two indicates that inhibitors of that target appear at twice the relative frequency among the antimalarial compounds than among the inactive ones, indicating that the target is more likely to be relevant for the antimalarial mode of action. Figure 2d shows the distribution of known target class activity when only human targets enriched by at least twofold, plus the antimicrobial targets, are considered. The added restriction reduced the total number of probable targets to 146 (from 413) and interestingly increased the relative proportion of kinases to 48%, strongly suggesting that this target class, still unexploited for antimalarials, can be a rich source of novel drug leads. Tables 1 and 2 show the number of

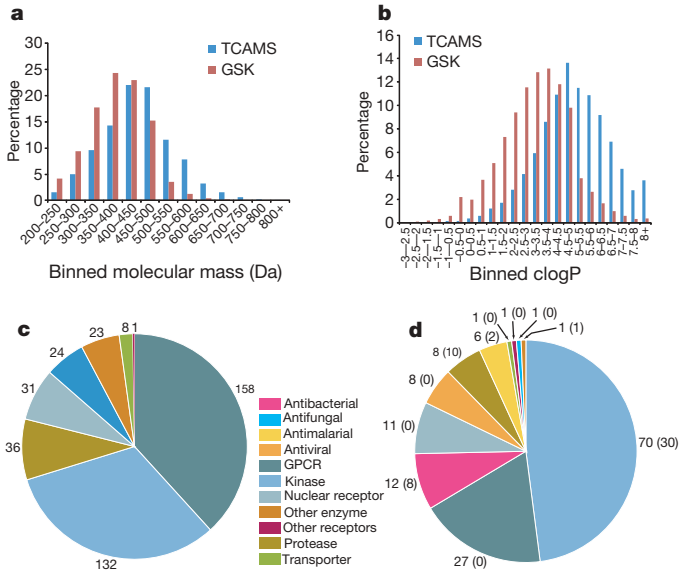


Figure 2 | Description of TCAMS and its target space. a, Relative molecular mass distributions of the TCAMS molecules and the general GSK screening collection. **b**, Relative clogP distributions of the TCAMS molecules and the general GSK screening collection. **c**, Distribution of target classes affected by the compounds in the set with potency and selectivity criteria described in the text. The number of targets in each class is indicated. **d**, Target classes remaining after applying the criterion that human targets be enriched at least twofold among the hits relative to the screening compound library, plus antimicrobial targets known to be relevant for *Plasmodium*. Number of targets in each class is indicated, with the number of identified malarial orthologues (BLASTP E-value $\leq 1.0^{-20}$) in parentheses.

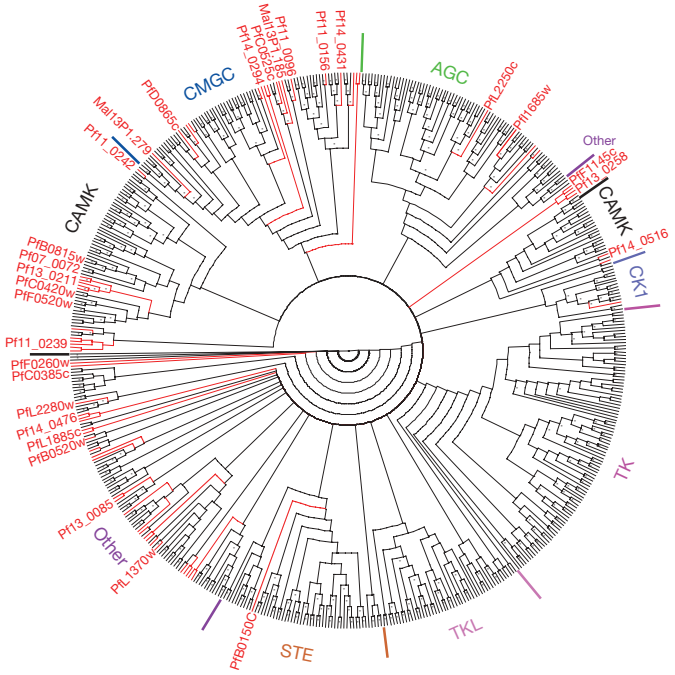
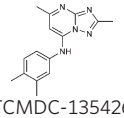
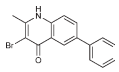
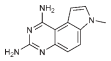
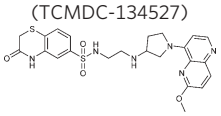
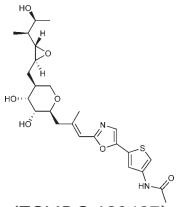
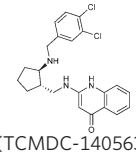
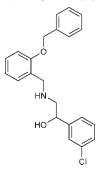
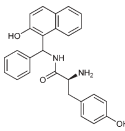
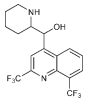
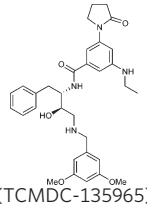
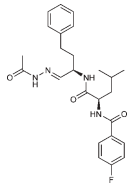
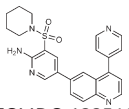
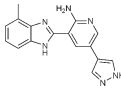
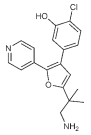
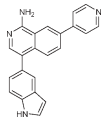
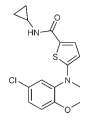


Figure 3 | Phylogenetic tree of combined human and *P. falciparum* kinomes. *P. falciparum* and human lineages are coloured red and black, respectively. Major human kinase subfamilies are labelled. TK: tyrosine kinases; TKL: tyrosine-like kinases; STE: homologues of yeast sterile 7, 11 and 20 kinases; CK1: casein kinase 1; AGC: PKA, PKG and PKC kinases; CAMK: calcium/calmodulin kinases; CMGC: CDK, MAPK, GSK3 and CLK kinases. Malarial kinases that are hypothesized targets in Table 1 are labelled. This neighbour-joining tree is based on pairwise amino acid sequence similarity for all known human[22] and *P. falciparum*[24,27] kinase domains. Nodes supported by $\geq 60\%$ of 1,000 bootstrap replicates are indicated by '*' (see Methods for tree reconstruction and Supplementary Fig. 1).

Table 1 | *P. falciparum* loci encoding proteins proposed as probable targets for TCAMS compounds

Target class	Target hypothesis/ *P. falciparum* locus	Number of chemotypes/number of inhibitors	Example inhibitor structures (compound ID)*
Antimalarial	Dihydroorotate dehydrogenase/ PFF0160c	5/20	(TCMDC-124418) 
	Electron transport chain†/PlfaoMp3	5/122	(TCMDC-135426) 
Antibacterial	Dihydrofolate reductase/ PFD0805w	4/9	(TCMDC-139863) 
	DNA gyrase/ PFL1120c PFL1915w	7/15	(TCMDC-134527) 
	Isoleucyl-tRNA synthetase/ PF13_0179	1/1	(TCMDC-131575) 
	Methionyl-tRNA synthetase/ PF10_0340	2/4	(TCMDC-139627) 
	Phenylalanyl-tRNA synthetase/ PFA0480w	3/7	(TCMDC-140563) 
	Tyrosyl-tRNA synthetase/ PF11_0181	2/2	(TCMDC-141232) 
	Ribosome complex†/ MAL1P3.03a	1/1	(TCMDC-123920) 
Proteases	Aspartic protease/ PF14_0076 PF14_0075 PF14_0077 PFC0495w	7/41	(TCMDC-134572) 
	Cysteine protease/ PF11_0162 PF11_0161 PF11_0165 PF11_0174 PF14_0553 PFL2290w	15/51	(TCMDC-135965) 
Kinases	Phosphatidylinositol 3-kinase/PFE0765w	12/16	(TCMDC-134265) 
	Ser/Thr protein kinase/PFL2250c PF11_0156 PF13_0085 PFI1685w PFF1145c PFD0865c PF14_0476 PF14_0431 PF11_0239 PF11_0096 PFC0525c MAL13P1.185 PF13_0258 MAL13P1.279 PF14_0294 PF14_0516 PFB0150c PFL2280w PFL1370w PFC0385c PFF0260w PFF0520w PF13_0211 PFC0420w PFL1885c PF11_0242 PFB0815w PF07_0072 PFB0520w	95/264	(TCMDC-133561)  (TCMDC-141334)  (TCMDC-133396) 
Other enzymes	β-Keto-acid reductase/ PFI1125c	2/2	(TCMDC-125270) 

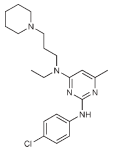
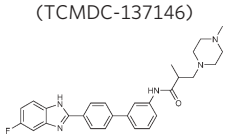
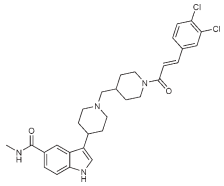
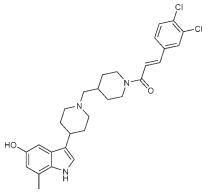
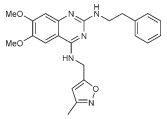
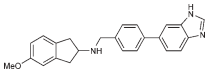
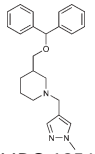
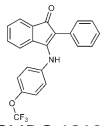
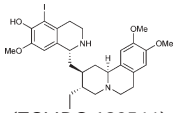
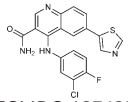
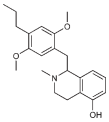
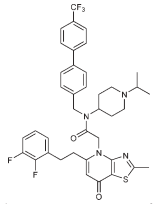
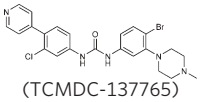
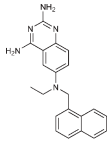
* Compound identification number as in Supplementary Table 1 (EXT_CMPD_NUMBER). Full compound assay data and structures are available at ChEMBL – Neglected Tropical Disease Archive (http://www.ebi.ac.uk/chemblntd).
† Multi-protein complexes.

chemotypes predicted to inhibit each target class. We have only added target annotations to compounds for which we had experimental evidence meeting the criteria above. Readers should use their own judgement when extrapolating proposed modes of action to chemical analogues.

We next searched for orthologues of the human targets in the *P. falciparum* genome, using conservative criteria (see Methods) for human–*P. falciparum* sequence homology to address this species' complex and divergent evolutionary relationship to mammals[16,17]. Forty-one *Plasmodium* proteins could be linked to human targets by those measures (Table 1). A number of compounds annotated as active in assays against other infectious agents, including *Plasmodium*, and with known mode of action, were also included in the analysis, yielding predictions of activity against organellar functions, such as protein synthesis, DNA metabolism or electron transport. This brings the total number of indicative *Plasmodium* targets to 51 (Table 1 and Supplementary Table 1).

To support the mode of action hypotheses we also checked gene expression profiles of the proposed targets using data from studies of the *P. falciparum* transcriptome across all life stages[18] and specifically, the intraerythrocytic development cycle measured over 48 h (ref. 19). Most targets proposed were expressed at various intervals in the intraerythrocytic malarial parasite and several were also present in other life stages (Supplementary Fig. 1).

Table 2 | Functions proposed as targets for TCAMS compounds without obvious homology to *P. falciparum* loci

Target class	Target hypothesis	Number of chemotypes/number of inhibitors	Example inhibitor structures* (compound ID)†
GPCR	Adrenergic receptor antagonist	6/9	(TCMDC-137488) 
	Cannabinoid receptor antagonist	1/1	
	Chemokine receptor antagonist	4/7	(TCMDC-137146) 
	Cholinergic receptor agonist	1/1	
	Free fatty acid receptor agonist	1/1	
	G protein-coupled receptor agonist	9/10	(TCMDC-139980) 
	G protein-coupled receptor antagonist	1/2	
	Serotonin receptor agonist	9/27	(TCMDC-140065) 
	Serotonin receptor antagonist	14/23	(TCMDC-135435) 
	Opioid receptor agonist	1/1	
	Opioid receptor antagonist	4/5	(TCMDC-133483) 
Ion channel	Ion channel inhibitor	71/108	(TCMDC-136042) 
Nuclear hormone receptor	Nuclear receptor agonist	7/9	(TCMDC-125489) 
	Nuclear receptor antagonist	9/14	(TCMDC-131829) 
	Peptide hormone receptor agonist	5/9	(TCMDC-138566) 
	Peptide hormone receptor antagonist	19/62	(TCMDC-137435) 
Other enzyme	Phospholipase inhibitor	3/3	(TCMDC-141264) 
	Lipid amide hydrolase inhibitor	3/3	(TCMDC-139221) 
	Serine protease	5/7	(TCMDC-137765) 
	Toll-like receptor agonist	1/1	

* Selected compounds where two or more chemotypes were identified.
† Compound identification number as in Supplementary Table 1 (EXT_CMPD_NUMBER). Full compound assay data and structures are available at ChEMBL – Neglected Tropical Disease Archive (http://www.ebi.ac.uk/chemblntd).

In these data, proteases and kinases are the most prominent human target classes with identified *P. falciparum* orthologues (Table 1 and Fig. 2d). Proteases are needed for bulk degradation of haemoglobin in the food vacuole, but also for signalling and remodelling of cellular structures, presenting a large array of potential lethal targets for TCAMS compounds[20]. But the larger proportion of hits with annotated activity against 'enriched' targets are protein kinase inhibitors, which provide new tools to exploit the malarial kinome for drug discovery[21]. The human genome encodes over 450 kinases[22] whereas *P. falciparum* has a much reduced complement of 85–99 kinases[23,24]. We reconstructed a phylogenetic tree of the entire complement of human and *P. falciparum* kinases to assist in assigning putative targets for kinase inhibitors. Figure 3 highlights the ancestral relationship of *Plasmodium* kinases to multiple human homologues, allowing inhibitors of several related human kinases to be mapped against a single malarial isoform, but keeping in mind that many kinase inhibitors have been shown to inhibit multiple, distantly related enzymes, often owing to three-dimensional structural similarities among binding sites[25].

Some compounds had historical assay activity that met our criteria for inclusion in the analysis, but against drug targets without obvious orthologues in the malarial genome, such as GPCRs, nuclear receptors, ion channels and transporters (Table 2). The simplest explanation would be that these compounds are killing *Plasmodium* through interactions with unrelated targets. More interestingly, *P. falciparum* could have essential proteins that are structurally and functionally similar to the human targets yet have no significant primary amino acid homology. This possibility is supported by recent analyses of structural rather than sequence homology, showing at least two candidate transmembrane proteins with similarity to human orphan GPCRs[26]. The screens in this study were conducted using *Plasmodium*-infected

erythrocytes; therefore, activity against specific human red blood cell targets is also a possible mode of action. Further exploration of this compound collection for disruptors of host–parasite interactions could lead to novel antimalarial therapeutic strategies, free from the selection pressures driving drug resistance in the parasite.

The public availability of this large set of potent and drug-like antiplasmodial compound structures enables important research questions to be addressed in future investigations, including characterizing the links between the biochemical activity of specific compounds and parasite killing and the extent of activity of these compounds in inhibiting *P. vivax*, which will be a critical attribute of new antimalarial treatments.

METHODS SUMMARY

Compounds were pre-dispensed in 384-well plates, RPMI/AlbuMAX growth media was added and *P. falciparum* inoculated in a biological safety level 3 laboratory, where plates were incubated for 72 h and then frozen at −70 °C overnight. LDH activity was quantified with the modified cofactor 3-acetylpyridine adenine dinucleotide (APAD) by measuring absorbance of the tetrazolium indicator nitro blue tetrazolium (NBT) at 650 nm, using an extensively modified LDH assay as described in Methods. Screening data were analysed using Microsoft Excel and Spotfire DecisionSite 8.2.1 software. Chemical structures were clustered using Daylight[13]. Mode of action hypotheses were generated searching archived target activity data of hits in the internal databases and filtering results for potency, selectivity, enrichment in the data set relative to the general screening collection and degree of homology to *P. falciparum* proteins, with the cut-off values described in the text. *Plasmodium* orthologues of human targets were searched and analysed using GCG Wisconsin Package v11.0 and PHYLIP 3.67. software packages as described in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 8 February; accepted 22 April 2010.

1. World Health Organization. World malaria report. ⟨http://www.who.int/malaria/publications/atoz/9789241563901/en/index.html⟩ (2009).
2. Anstey, N. M., Russell, B., Yeo, T. W. & Price, R. N. The pathophysiology of vivax malaria. *Trends Parasitol.* **25,** 220–227 (2009).
3. Ekland, E. H. & Fidock, D. A. *In vitro* evaluations of antimalarial drugs and their relevance to clinical outcomes. *Int. J. Parasitol.* **38,** 743–747 (2008).
4. Andriantsoanirina, V. *et al. Plasmodium falciparum* drug resistance in Madagascar: facing the spread of unusual *pfdhfr* and *pfmdr-1* haplotypes and the decrease of dihydroartemisinin susceptibility. *Antimicrob. Agents Chemother.* **53,** 4588–4597 (2009).
5. Bonnet, M. *et al.* Varying efficacy of artesunate+amodiaquine and artesunate+sulphadoxine-pyrimethamine for the treatment of uncomplicated falciparum malaria in the Democratic Republic of Congo: a report of two in-vivo studies. *Malar. J.* **8,** 192 (2009).
6. Carrara, V. I. *et al.* Changes in the treatment responses to artesunate-mefloquine on the northwestern border of Thailand during 13 years of continuous deployment. *PLoS ONE* **4,** e4551 (2009).
7. Payne, D. J., Gwynn, M. N., Holmes, D. J. & Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nature Rev. Drug Discov.* **6,** 29–40 (2007).
8. O'Brien, P. J. *et al.* High concordance of drug-induced human hepatotoxicity with *in vitro* cytotoxicity measured in a novel cell-based model using high content screening. *Arch. Toxicol.* **80,** 580–604 (2006).
9. Xia, M. *et al.* Compound cytotoxicity profiling using quantitative high-throughput screening. *Environ. Health Perspect.* **116,** 284–291 (2008).
10. Yuan, J. *et al.* Genetic mapping of targets mediating differential chemical phenotypes in *Plasmodium falciparum*. *Nature Chem. Biol.* **5,** 765–771 (2009).
11. Reed, M. B., Saliba, K. J., Caruana, S. R., Kirk, K. & Cowman, A. F. Pgh1 modulates sensitivity and resistance to multiple antimalarials in *Plasmodium falciparum*. *Nature* **403,** 906–909 (2000).
12. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39,** 2887–2893 (1996).
13. Daylight Chemical Information Systems, Inc. Daylight theory manual. ⟨http://www.daylight.com/dayhtml/doc/theory/index.html⟩ (2008).
14. Johnson, M., Lajiness, M. & Maggiora, G. M. in *Qsar: Quantitative structure-activity relationships in drug design* (ed Fauchere, J. L.) 167–171 (Alan R. Liss, 1989).
15. Davis, A. M., Keeling, D. J., Steele, J., Tomkinson, N. P. & Tinker, A. C. Components of successful lead generation. *Curr. Top. Med. Chem.* **5,** 421–439 (2005).
16. Köhler, S. *et al.* A plastid of probable green algal origin in apicomplexan parasites. *Science* **275,** 1485–1489 (1997).
17. McFadden, G. I., Reith, M. E., Munholland, J. & Lang-Unnasch, N. Plastid in human parasites. *Nature* **381,** 482 (1996).
18. Le Roch, K. G. *et al.* Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301,** 1503–1508 (2003).
19. Bozdech, Z. *et al.* The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.* **1,** e5 (2003).
20. Wu, Y., Wang, X., Liu, X. & Wang, Y. Data-mining approaches reveal hidden families of proteases in the genome of malaria parasite. *Genome Res.* **13,** 601–616 (2003).
21. Leroy, D. & Doerig, C. Drugging the *Plasmodium* kinome: The benefits of academia–industry synergy. *Trends Pharmacol. Sci.* **29,** 241–249 (2008).
22. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298,** 1912–1934 (2002).
23. Anamika, S. N. & Krupa, A. A genomic perspective of protein kinases in *Plasmodium falciparum*. *Proteins* **58,** 180–189 (2005).
24. Ward, P., Equinet, L., Packer, J. & Doerig, C. Protein kinases of the human malaria parasite *Plasmodium falciparum*: the kinome of a divergent eukaryote. *BMC Genomics* **5,** 79 (2004).
25. Karaman, M. W. *et al.* A quantitative analysis of kinase inhibitor selectivity. *Nature Biotechnol.* **26,** 127–132 (2008).
26. Madeira, L. *et al.* Genome-wide detection of serpentine receptor-like proteins in malaria parasites. *PLoS ONE* **3,** e1889 (2008).
27. Aurrecoechea, C. *et al.* PlasmoDB: A functional genomic database for malaria parasites. *Nucleic Acids Res.* **37,** D539–D543 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Peregrina and S. Prats for technical assistance, and D. Jiménez-Alfaro for supplying compounds pre-dispensed in microtitre plates. We thank P. Vallance and R. Keenan for organising support for this work, R. Macarron and J. Luengo for developing the IFI index and getting the chemistry data ready for publication, together with J. M. Fiandor, S. Chakravorty and members of GSK's Chemistry Council. J. Lewis, A. Clow, J. Overington and M. Davies were instrumental in the uploading and formatting of the data. We also thank N. Cammack, P. Sanseau and J. Burrows for critically commenting on the manuscript. The support and funding of Medicines for Malaria Venture is gratefully acknowledged.

Author Contributions F.-J.G. and J.F.G.-B. planned and designed the work. F.-J.G. supervised all experimental work and analysed the screening data, L.M.S., J.V., C.d.C. and E.A. performed the screening assays and contributed to data analysis. J.-L.L., D.E.V., D.V.S.G. and C.E.P. performed the cheminformatic analyses and wrote sections of the manuscript. V.K., S.H. and J.R.B. performed the bioinformatic analyses and J.R.B. contributed the relevant sections to the manuscript. L.R.C and J.F.G.-B integrated individual contributions and wrote the final manuscript.

Author Information Chemical structures and data described have been deposited at EBI (http://www.ebi.ac.uk/chemblntd). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.F.G.-B. (jose.f.garcia-bustos@gsk.com).

METHODS

Parasites. *P. falciparum* strains 3D7 and Dd2 used in this study were obtained from the Malaria Research and Reference Reagent Resource Center (MR4). Accurate descriptions can be obtained at http://www.mr4.org. Parasite strains were cultured using standard procedures as described[28]. An inoculum of parasitized red blood cells (PRBC) at 0.25% parasitaemia and 2% haematocrit in RPMI-1640, 5% AlbuMAX, 2% D-sucrose, 0.3% glutamine and 150 μM hypoxanthine was used for the assay.

Growth inhibition assay. Assay plates (384-well) were prepared by dispensing 0.05 μl of compound from master plates at 1 mM in each well. Final assay volume was 25 μl and final compound concentration was 2 μM. The sixth column was the positive growth control and had 0.05 μl of DMSO. The eighteenth column had 0.05 μl of a mixture of 50 μM chloroquine and 50 μM artemisinin stock solutions as negative growth control. The parasite inoculum (25 μl) was dispensed into plates containing compounds using a Multidrop Combi dispenser (Thermo Scientific). Plates were shaken for 10 s to ensure mixing and then incubated at 37 °C for 72 h in an atmosphere of 5% CO_2, 5% O_2, 95% N_2.

Evaluation of parasite growth using LDH activity. After 72 h of incubation, plates were frozen at −70 °C overnight and then thawed at room temperature for at least 4 h. To evaluate LDH activity, 70 μl of freshly made reaction mix (143 mM sodium L-lactate, 143 μM 3-acetyl pyridine adenine dinucleotide (APAD), 178.75 μM Nitro Blue tetrazolium chloride (NBT), 286 μg ml^{-1} diaphorase (2.83 U ml^{-1}), 0.7% Tween 20, 100 mM Tris-HCl pH 8.0) was dispensed using a Multidrop Combi dispenser. Plates were shaken to ensure mixing and absorbance at 650 nm was monitored in a plate reader after 10 min of incubation at room temperature. Data were normalized to percent growth inhibition using positive and negative controls and the equation:

$$\text{Percentage inhibition growth} = \left[1 - \left(\frac{A_{\text{well}} - A_{\text{neg}}}{A_{\text{pos}} - A_{\text{neg}}}\right)\right] \times 100$$

where A_{well} is the absorbance measured in the well, and A_{pos} and A_{neg} are the average absorbances measured for the positive and negative controls, respectively. This method is a modification of existing ones[29] that requires only a single pipetting step after compound incubation and gave a signal to noise ratio of 10 under the conditions chosen. The approach allowed kinetic and end-point readouts and produced a Z′ quality factor[30] higher than 0.7 in validation assays (Supplementary Fig. 2). Potencies of standard antimalarial agents in this assay were comparable to those determined by the current gold-standard, 96-well, hypoxanthine incorporation assay[31] (Supplementary Table 3).

At this level of miniaturization, integrity of erythrocytes and LDH activity can be inspected visually, allowing for rapid detection of dispensing errors, interference by coloured compounds, or haemolysis, making the method very useful for low technology settings (Supplementary Fig. 3). Proliferation of asynchronous parasites was measured after 72 h of incubation in the presence of 2 μM compound. We chose a 72 h incubation time to ensure all parasites traversed at least once through each stage of the cell cycle and to increase the chances of identifying slow acting and 'delayed death phenotype' inhibitors[32,33]. Some of these were indeed detected (for example, tetracyclines), but others, presumably with a more delayed effect, were not (for example, azythromycin). Metrics for the full screen were of high quality, with an average Z′ of 0.7 and fewer than 3,000 wells lost to analysis (Supplementary Table 4).

Given the large number of positives, it was operationally necessary to estimate the concentrations producing 50% inhibition using the LDH assay above and generating dose–response curves with fivefold dilution steps down to 3 nM compound in an interplate design, instead of using the hypoxanthine incorporation assay with twofold dilution intraplate series generally considered the standard method to calculate IC_{50} for antimalarials[34]. We denominated this parameter XC_{50} to indicate that it is an estimation of the usual IC_{50} values. The lowest concentration tested was 3 nM. Agreement between the two methods was found to be within the expected limits with standard antimalarials (Supplementary Table 3). To eliminate the possibility of retaining inhibitors of the biochemical readout system, one set of the primary hits was assayed against parasite LDH activity under identical screening conditions. Non-specific cytotoxicity was also a concern and we addressed it by assaying the hits at five times the screening concentration against human hepatoma HepG2 cells as below.

Preparation of extracts to evaluate direct LDH inhibition by hit compounds. *P. falciparum* 3D7 strain was grown as described above at 37 °C for 72 h. The culture was frozen at −70 °C overnight. Cultures were thawed at room temperature for at least 4 h and used as crude extracts (25 μl per well) to measure the possible direct inhibition of LDH by the compounds, assayed as above.

Cytotoxicity assay. Actively growing human hepatoma HepG2 cells were removed from a T-175 TC flask using Cell Dispersion Medium (5 ml), and dispersed by repeated pipetting. Cell suspensions were added to 500 ml of medium (Eagle's minimum essential medium with Earle's salts and L-glutamine, 10% FBS, 1% non-essential amino acid solution, 1% penicillin and streptomycin). Twenty five microlitres were dispensed into wells with compounds in Greiner TC-treated, black 384-well clear bottom plates using a Multidrop dispenser. Seeding density was adjusted to ensure that new monolayers were not more than ~ 50% confluent (typically 3,000 cells per well). Cells were left at 37 °C, 5% CO_2 in a humidified incubator in the presence of 10 μM compound for 48 h. Intracellular ATP was evaluated using a luciferase-coupled ATP quantification assay (CellTiter-Glo, Promega) according to the manufacturer's instructions.

Computational analysis. Molecular frameworks were calculated using an in-house implementation of the algorithm described previously[12]. The frameworks were then used to define clusters of compounds that share a particular framework. To minimise the number of frameworks that describe only a small number of molecules (<10 in our standard processing), we attempt to reclassify all such molecules into a smaller framework (that is, the framework describes a smaller proportion of the structure of the reclassified molecule). Therefore there will exist some clusters that share a molecular framework, but that are not as structurally homogenous as would be given by simply grouping together compounds by unique frameworks. Compounds were also classified in chemical families using the Daylight fingerprint methods with a Tanimoto similarity index of 0.85, following the procedures in the Daylight Information Systems manual[13].

Among all compounds tested in the antimalarial assay, ~130,000 had proprietary assay data against one or more human proteins from previous drug development screening campaigns, including 3,435 of the active compounds, which were associated to 413 human proteins with potencies matching the inclusion criteria described in the main text. A further 770 compounds were associated with specific programmes involving bacterial, malarial or antiviral targets. To exclude promiscuous and non-specific compounds from the analysis, an 'inhibition frequency index' (IFI) was calculated. This is the relative frequency with which a compound has scored more than 50% inhibition in an HTS assay, calculated with the equation:

$$\text{IFI} = \frac{\text{Number of HTS assays where \% inhibition} \geq 50\%}{\text{Total number of HTS assays}} \times 100,$$

excluding kinase assays. For example, if a compound was screened in 100 HTS assays, and showed ≥ 50% inhibition in 20 of those, the IFI = 20. The IFI threshold used to exclude compounds was based on the total number of screens in which they have been tested, ranging from 5% for compounds tested in > 100 assays, to 20% for compounds tested in > 25 assays. The threshold above which compound efficacy against specific human targets was considered significant was defined as $pIC50 \geq 7$ for inhibition or antagonist assays, and $pEC50 \geq 6.5$, for agonist, activation, or modulator assays.

Activities at more than 600 target-result type combinations, (some targets are assayed in both an antagonist and agonist mode) were analysed amongst the antimalarial compounds, representing potential modes of action. The target activities for the screened compounds were analysed to identify targets over-represented amongst the antimalarial actives versus inactives. An 'enrichment' was calculated for each possible target-result type combination by dividing the hit rate for that target amongst antimalarial compounds, by the hit rate at that target for all compounds screened in the assay:

$$\text{Enrichment} = \frac{x/n}{X/N}$$

where x, the number of antimalarial hits with an activity at the target above the potency threshold; n, the number of the antimalarial hits with a measured pIC_{50} or pEC_{50} at the target; X, the number of compounds from the entire screening set with an activity at the target above the potency threshold; N, the number of compounds from the entire screening set with a measured pIC_{50} or pEC_{50} at the target.

Compounds originating from programmes in infectious diseases were not well represented in the historical target assay data. These compounds were annotated with their originating target and therapeutic class (antimalarial, antibacterial, antiviral, antifungal) and included in the counts and target classes illustrated in Fig. 2d to provide a more complete view of potential mechanisms represented among the TCAMS actives.

Literature compounds and target activities in Supplementary Table 2 were retrieved from the AURQUEST database[35]. TCAMS actives were used to search for similar compounds in the database, as defined by Daylight fingerprints and a Tanimoto similarity coefficient of 0.85 or higher[13]. SciFinder (version 2007.2, American Chemical Society) was used to search for compounds active at protein targets not represented in AURQUEST. As some molecules, particularly kinase inhibitors, have a very large number of close analogues in the public domain, the most similar 100 compounds were kept. The biological results were restricted to those data with a dose–response curve, and a pXC_{50} of 6 or higher.

Using BLASTP[36] we queried the protein complement of *P. falciparum* 3D7 genome with RefSeq proteins for all human, bacterial and viral targets accepting a homology cut-off of an E-value $\leq 1.0 \times 10^{-20}$. The top three hits were collected for most proteins, except for kinases where more detailed phylogenetic analyses were performed (see below). Human–malarial homology was confirmed by reciprocal BLASTP searches of identified *P. falciparum* homologues against the human RefSeq protein databases. *P. falciparum* gene locus identifiers and annotations were obtained from PlasmoDB[27].

A phylogenetic tree of the combined *P. falciparum* and *Homo sapiens* kinomes was reconstructed using all annotated human and *P. falciparum* kinases as annotated by Interpro kinase queries of PlasmoDB, as well as from a previously reported *P. falciparum* kinome[24]. Initial multiple sequence alignments were performed using the program CLUSTALW v1.8 (ref. 37) with default settings and subsequently refined manually using the program SEQLAB of the GCG Wisconsin Package v11.0 software package (Accelrys). Redundant *P. falciparum* kinases were removed after an initial multiple sequence alignment. Residues that could not be unambiguously aligned or that contained insertions or deletions were removed from the multiple sequence alignment before phylogenetic analysis. The final kinase multiple sequence alignment was 160 amino acids in length and included a combined total of 557 human and malarial kinases (available as a Supplementary file).

We constructed phylogenetic trees using distance neighbour-joining (NJ) method. NJ trees were based on pair-wise distances between amino acid sequences using the programs NEIGHBOR and PROTDIST (Dayhoff option) of the PHYLIP 3.67 package[38]. The programs SEQBOOT and CONSENSE were used to estimate the confidence limits of branching points from 1,000 bootstrap replications. All trees were visualized using the program DENDROSCOPE[39] (Fig. 3).

28. Trager, W. & Jensen, J. B. Human malaria parasites in continuous culture. *Science* **193,** 673–675 (1976).
29. Makler, M. T. & Hinrichs, D. J. Measurement of the lactate dehydrogenase activity of *Plasmodium falciparum* as an assessment of parasitemia. *Am. J. Trop. Med. Hyg.* **48,** 205–210 (1993).
30. Zhang, J. H., Chung, T. D. Y. & Oldenburg, K. R. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.* **4,** 67–73 (1999).
31. Desjardins, R. E., Canfield, C. J., Haynes, J. D. & Chulay, J. D. Quantitative assessment of antimalarial activity in vitro by a semiautomated microdilution technique. *Antimicrob. Agents Chemother.* **16,** 710–718 (1979).
32. Goodman, C. D., Su, V. & McFadden, G. I. The effects of anti-bacterials on the malaria parasite *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **152,** 181–191 (2007).
33. Ramya, T. N. C., Mishra, S., Karmodiya, K., Surolia, N. & Surolia, A. Inhibitors of nonhousekeeping functions of the apicoplast defy delayed death in *Plasmodium falciparum*. *Antimicrob. Agents Chemother.* **51,** 307–316 (2007).
34. Fidock, D. A., Rosenthal, P. J., Croft, S. L., Brun, R. & Nwaka, S. Antimalarial drug discovery: Efficacy models for compound screening. *Nature Rev. Drug Discov.* **3,** 509–520 (2004).
35. Aureus Pharma AurSCOPE database. ⟨http://www.aureus-pharma.com/Pages/Products/Aurscope.php⟩ (2004).
36. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
37. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
38. Felsentein, J. Phylip (phylogenetic inference package) v3.67. *Department of Genetics, University of Washington* ⟨http://evolution.genetics.washington.edu/phylip.html⟩.
39. Huson, D. H. *et al.* Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8,** 460 (2007).