

Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries

Henrike Veith¹, Noel Southall¹, Ruili Huang¹, Tim James², Darren Fayne², Natalia Artemenko², Min Shen¹, James Inglese¹, Christopher P Austin¹, David G Lloyd² & Douglas S Auld¹

The cytochrome P450 (CYP) gene family catalyzes drug metabolism and bioactivation and is therefore relevant to drug development. We determined potency values for 17,143 compounds against five recombinant CYP isozymes (1A2, 2C9, 2C19, 2D6 and 3A4) using an *in vitro* bioluminescent assay. The compounds included libraries of US Food and Drug Administration (FDA)-approved drugs and screening libraries. We observed cross-library isozyme inhibition (30–78%) with important differences between libraries. Whereas only 7% of the typical screening library was inactive against all five isozymes, 33% of FDA-approved drugs were inactive, reflecting the optimized pharmacological properties of the latter. Our results suggest that low CYP 2C isozyme activity is a common property of drugs, whereas other isozymes, such as CYP 2D6, show little discrimination between drugs and unoptimized compounds found in screening libraries. We also identified chemical substructures that differentiated between the five isozymes. The pharmacological compendium described here should further the understanding of CYP isozymes.

The human cytochrome P450 family contains 57 isozymes, which function in normal metabolism, influence drug pharmacokinetics and effect negative outcomes in patients through drug-drug interactions^{1,2}. The CYP isozymes metabolize approximately two-thirds of known drugs in humans, with 80% of this attributable to five isozymes—1A2, 2C9, 2C19, 2D6 and 3A4 (ref. 3).

There has been increased effort to minimize liabilities due to CYP isozyme activity in drug discovery by incorporating early-stage *in vitro* metabolic characterization of compounds⁴. Interactions between a limited number of compounds (typically, <100) and CYP isozymes have been described^{5–8}, but results from these studies are difficult to compare because of differences in technologies, assay conditions and data analysis methods⁹. Despite the high interest in this gene family, few public databases exist (e.g., <http://www.bindingdb.org/>), and the scientific literature remains fragmented, making data mining difficult.

A public CYP bioactivity database would provide a foundation for testing and improving current CYP activity prediction models and for guiding the use of *in vitro* CYP assays in early-phase drug discovery efforts. We have previously shown that the enzyme activity profiles needed for such a database can be obtained using quantitative high-throughput screening (qHTS) with assays using either purified enzymes^{10–12} or complex cell-based activity profiles^{13,14}. In qHTS, libraries of small molecules are assayed at multiple concentrations, producing concentration-response curves (CRCs) and potencies for every compound¹⁰. We describe here the application of qHTS to construct a CYP bioactivity database that should provide a foundation for testing and improving current CYP activity prediction models as well guide the use of *in vitro* CYP assays in early-phase drug discovery efforts.

RESULTS

qHTS of CYP isozymes

Using qHTS with a bioluminescent assay¹⁵, we screened 17,143 chemical compounds at 7–15 concentrations against five CYP isozymes (CYP 1A2, 2C9, 2C19, 2D6 and 3A4). In the assay, active CYP enzyme converts proluciferin substrates to luciferin, which is then utilized by firefly luciferase to generate a luminescent signal. We screened (i) 8,019 compounds from the Molecular Libraries Small Molecule Repository (MLSMR; <http://www.ncbi.nlm.nih.gov/sites/entrez?db=pcsubstance&term=mlsmr>), including compounds chosen for diversity and rule-of-five compliance¹⁶, synthetic tractability and availability; (ii) 6,144 compounds from biofocused libraries, which included 1,114 FDA-approved drugs; and (iii) 2,980 compounds from combinatorial chemistry libraries, containing privileged structures targeted at G protein-coupled receptors and kinases and containing purified natural products or related structures. As is typical in qHTS, we fit the Hill equation to the data to generate CRCs for every compound tested, using a method previously described¹⁰.

CRCs were divided into five categories based on the potency, efficacy and quality of the curve that was fit to the observed response: inhibitory CRCs of high confidence (category 1) and low confidence (category 2), activator CRCs of high confidence (category 3) and low confidence (category 4), and inactive (category 5), if no response was observed up to the highest tested compound concentration (57 μ M). All five isozyme assays showed Z' -factors averaging ~ 0.6 , indicating good performance. In retesting a set of 91 randomly selected compounds, we observed excellent confirmation of activity (84–90%) for all five isozymes (Online Methods and **Supplementary Fig. 1**). We note that an inhibitory CRC in the present data set could result from a compound acting as a bona fide inhibitor or as a substrate because

¹NIH Chemical Genomics Center, National Institutes of Health, Bethesda, Maryland, USA. ²Molecular Design Group, School of Biochemistry and Immunology, Trinity College, Dublin, Ireland. Correspondence should be addressed to D.S.A. (dauld@mail.nih.gov).

Received 21 August; accepted 23 September; published online XX XX 2009; doi:10.1038/nbt.XXX

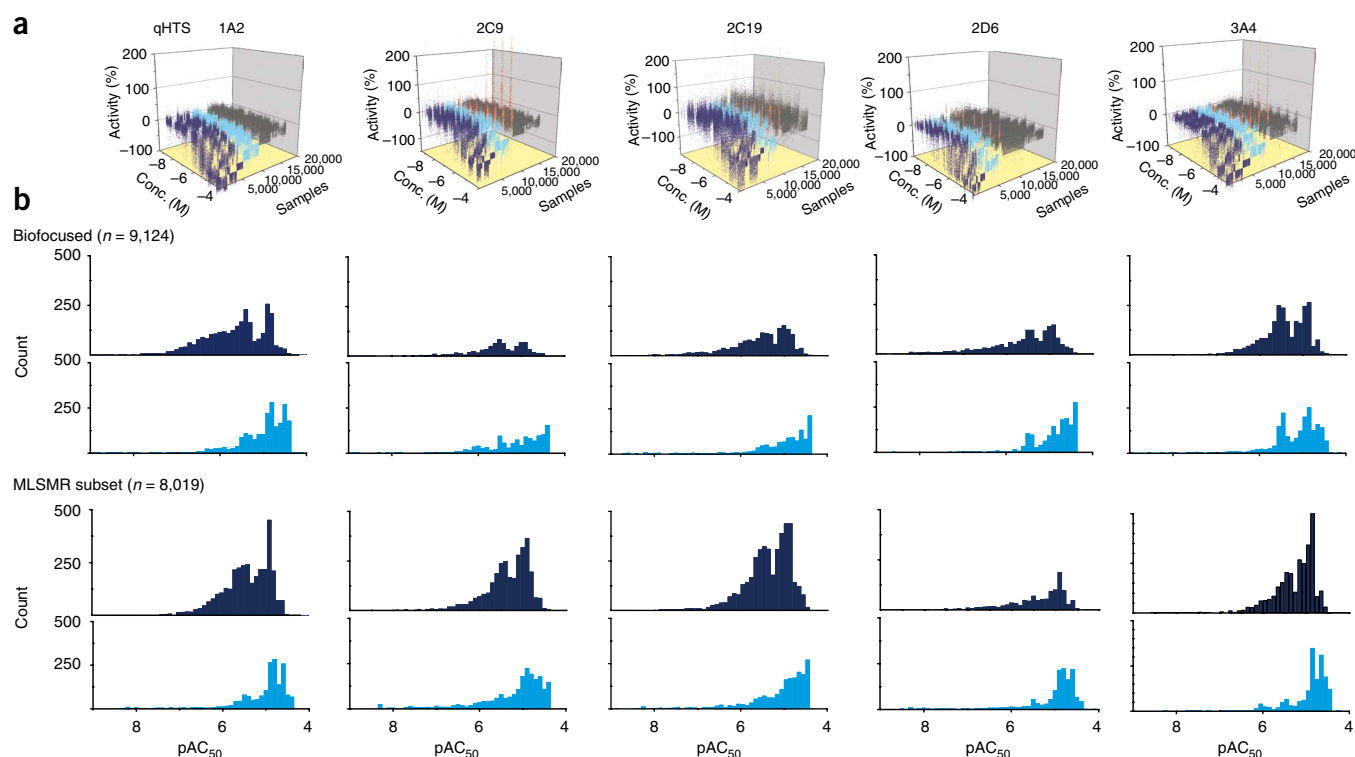


Figure 1 qHTS of five cytochrome P450 isoforms (CYP 1A2, 2C9, 2C19, 2D6 and 3A4). **(a)** Activity of isoforms in the presence of ~17,000 compounds at various concentrations. Dark blue or red represent data that fit to high-confidence CRCs (inhibitors or activators, respectively). Light blue or red are data that fit to low-confidence CRCs (inhibitors or activators, respectively). Inactive compounds that did not show any concentration response are gray. Conc., log₁₀ concentration of compound. **(b)** Potency distribution for high-confidence (dark blue) and low confidence (light blue) inhibitory CRCs for the biofocused (including the FDA-approved drugs) and MLSMR sets. Scales are identical for both upper and lower graphs and are shown at left and bottom of the graphs.

both will compete for free CYP enzyme (**Supplementary Fig. 2**). The qHTS data for each of the five CYPs are shown in **Figure 1a** and are available in PubChem.

Analysis and comparison of CYP activity

A high degree of inhibitory activity for all five isoforms was observed in the 17,000-compound collection (30–78% of the collection inhibited any one CYP isozyme). High-confidence activating CRCs (compounds that increased the rate of proluciferin conversion) were only appreciably observed in CYP 3A4 (2.5%) and CYP 2C9/2C19 (~3.5% each). Activation of CYP isozyme activity is typically substrate dependent^{17,18}, so the present assays would not be expected to comprehensively characterize this type of behavior.

By examining the distribution of half-maximal inhibitory concentration (IC₅₀) within high- and low-confidence inhibitory compounds (**Fig. 1b**), we observed that compounds in the biofocused and MLSMR libraries were selective for specific CYP isoforms. For example, both libraries showed a similar frequency of activity towards CYP 2D6, whereas the biofocused library was considerably less active against CYP 2C9 and 2C19 (**Fig. 1b**) and to a lesser extent against 1A2 and 3A4. When the distribution of potency between the MLSMR and FDA-approved drugs was compared, we found few significant isozyme-specific differences except for CYP 1A2, where the FDA drugs appeared to be less potent than the MLSMR, and CYP 2D6, where the FDA drugs were more potent than the MLSMR ($P < 0.01$; **Supplementary Fig. 3**). We observed additional differences between the biofocused, MLSMR and combinatorial chemistry compounds in both the number and distribution of inhibitory CRCs

across the five CYP isoforms. In the MLSMR an average of $58 \pm 16\%$ of compounds were found to be active against any specific isozyme. The biofocused library showed approximately half this activity (average of $32 \pm 8\%$). FDA drugs (18% of the biofocused library) showed similar activity (average of $31 \pm 8\%$). In contrast, the combinatorial library showed an average activity ($52 \pm 19\%$) similar to that of the MLSMR subset. CYP 1A2 and CYP 3A4 showed strong activity for a specific class of quinazoline compounds (60–75% of this compound class was active), which is a common scaffold found among protein kinase inhibitors.

The importance of FDA-approved drugs, and the large difference in activity between these and the MLSMR, prompted us to further compare the CYP activity between the two collections. As shown in **Figure 2a**, 33.3% of FDA-approved drugs were inactive against all five isoforms compared to 7.1% of compounds from the MLSMR subset. Pan-activity (e.g., inhibition against all five isoforms) was increased approximately twofold in the MLSMR subset compared to the FDA drugs (8.0% and 3.8%, respectively). Also, there was a steady decline in the percentage of active compounds when more than one CYP was active in the FDA-approved drugs (**Fig. 2a**). FDA drugs were less active, both in terms of the percentage of compounds interacting with any isozyme combination, and the average number of isoforms interacting with each compound. Therefore, historically drugs have been optimized to have low combination CYP activity, with fewer drugs showing activity against more than three CYPs.

The two CYP 2C isoforms (2C9 and CYP 2C19) were the least affected by the FDA drugs compared to MLSMR compounds (46% and 57% fewer compounds in the FDA drugs, respectively;

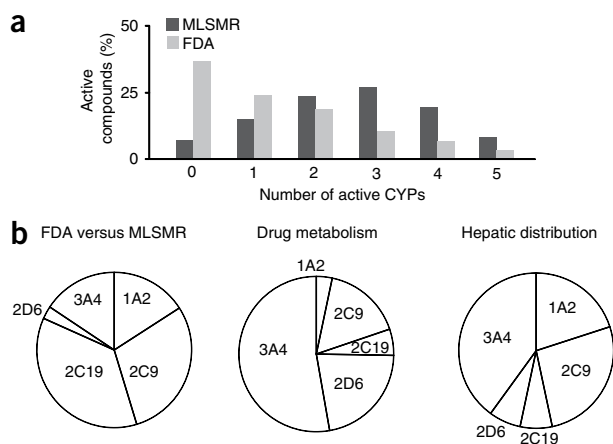


Fig. 2b). Both CYP 3A4 and 1A2 were differentially affected to a lesser degree (~24% difference for each), and CYP 2D6 showed a 5% difference between the two libraries. We also found that when only one isozyme was active this was unlikely to be either CYP 2C19 or 2C9 for the MLSMR (**Supplementary Fig. 4**). Therefore, although hepatic CYP activity and distribution (**Fig. 2b**) emphasizes the role of CYP 3A4 in drug metabolism, drugs appear to have been optimized to have low CYP 2C9 and 2C19 inhibitory activity.

Identification of active chemical series

To assess whether structurally similar compounds affected specific CYP isozymes, we clustered all 17,000 compounds based on their structural similarity and represented these as self-organizing maps (SOMs; **Fig. 3**)^{19,20}. In the SOMs, each hexagon represents a cluster of structurally similar compounds, with neighboring hexagons containing more similar structures than distal hexagons. This approach identified a quinazoline class of compounds that affected CYP 1A2 but not the other four isozymes (**Fig. 3**, hexagons along the lower edge that are red in the CYP 1A2 SOM and blue in others). We also found structures that affected combinations of isozymes: two hexagons in the bottom left corner in the SOMs represent structures that are active (red) against CYP 2C19, 2C9 and 3A4 but inactive (blue) against CYP 1A2 and 2D6. The number of compounds active against all isozymes is relatively small ($n = 350$), and all share a relatively high degree of structural similarity, reflected in the few hexagons that are clustered together and colored red in all five SOMs. Conversely, the number of pan-inactive compounds is quite large ($n > 2,000$, blue cluster of hexagons in all five SOMs).

Figure 3 Clustering of CYP isozyme activity across the 17,000-compound collection. SOMs are shown in which each hexagon represents a cluster of compounds showing structural similarity. Red clusters represent structures enriched in compounds active against the CYP enzyme as measured by a Fisher's exact test. Blue clusters represent structures rarely found in active compounds. A darker red or blue color indicates a higher level of enrichment or deficiency, respectively, in active compounds. Hexagons are colored by the negative $\log_{10}$ of the P -values (not corrected for multiple hypothesis testing). $P < 0.05$ is considered statistically significant. For example, the group of blue hexagons that consistently appear in the top middle region of the SOMs represent structurally related compounds that tend to be inactive against all five CYP isozymes. Isozyme SOMs are ordered left to right according to hierarchical clustering using Minkowski distance.

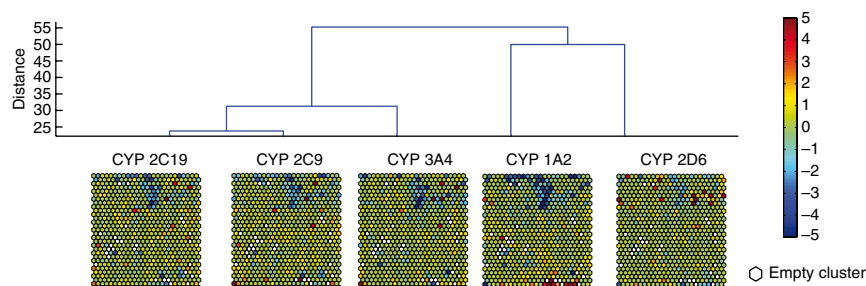


Figure 2 Distribution and differences in CYP activity between MLSMR versus FDA sets and comparison to published descriptions. (a) Distribution of MLSMR (dark gray) and FDA (light gray) compounds that were active against a given number of CYP isozymes. (b) Difference in the percentage of active compounds in the FDA and MLSMR libraries (left pie chart). Distribution of hepatic CYP metabolic activity (middle pie chart) and expression of each isozyme (right pie chart). Hepatic metabolism and expression data are from ref. 26.

Once the activity data were organized by SOMs, we could relate the activity patterns to the genetic sequence similarity of the human CYP isozymes. We hierarchically clustered (using a Minkowski distance as the similarity metric) the five isozymes using the compound activity patterns, and the resulting dendrogram is shown in **Figure 3**. Clustering divided the five isozymes into two major groups, one consisted of CYP 2C19, 2C9 and 3A4, with CYP 2C9 and 2C19, which are the most similar in sequence, having the most similar activity patterns, and the other one consisted of CYP 1A2 and 2D6, although these latter two showed a lesser degree of activity similarity.

Identification of activity-differentiating substructures

To identify structural features that differentiate between active and inactive compounds, we searched for substructures disproportionately represented in particular CRC classes relative to the entire testing set (16,144 unique compounds from the 17,000-compound library; see Online Methods). A selection of the results for structural features that differentiated between active and inactive compounds or showed isoform specific behavior is given in **Figure 4** (additional substructures are in **Supplementary Table 1**). To obtain the most accurate associations, we focused on the high-confidence inhibitory CRC classes and the inactive class. Activating CRC categories are not shown because of the relative dearth of records in this class.

A number of chemical substructures were associated with inhibitory and inactive compounds (**Fig. 4**). The presence of an aliphatic alcohol group (1) was associated with compounds inactive against four of the five isozymes. The presence of an aromatic hydroxyl group was associated with a similar but weaker trend, as was the presence of an ether linkage (data not shown). The presence of a primary aliphatic amine (2) or a quaternary ammonium salt (5) was also associated with compounds showing pan-isoform inactivity. In contrast, secondary and tertiary aliphatic amines (3 and 4, respectively) were associated with isoform-specific behavior, associated with the inhibitor and/or substrate response class for CYP 2D6 but with the inactive class for other isozymes. This is consistent with the known preference of CYP 2D6 for substrates containing basic, protonatable nitrogen atoms²¹.

Q13

Q2

Q3

Q4

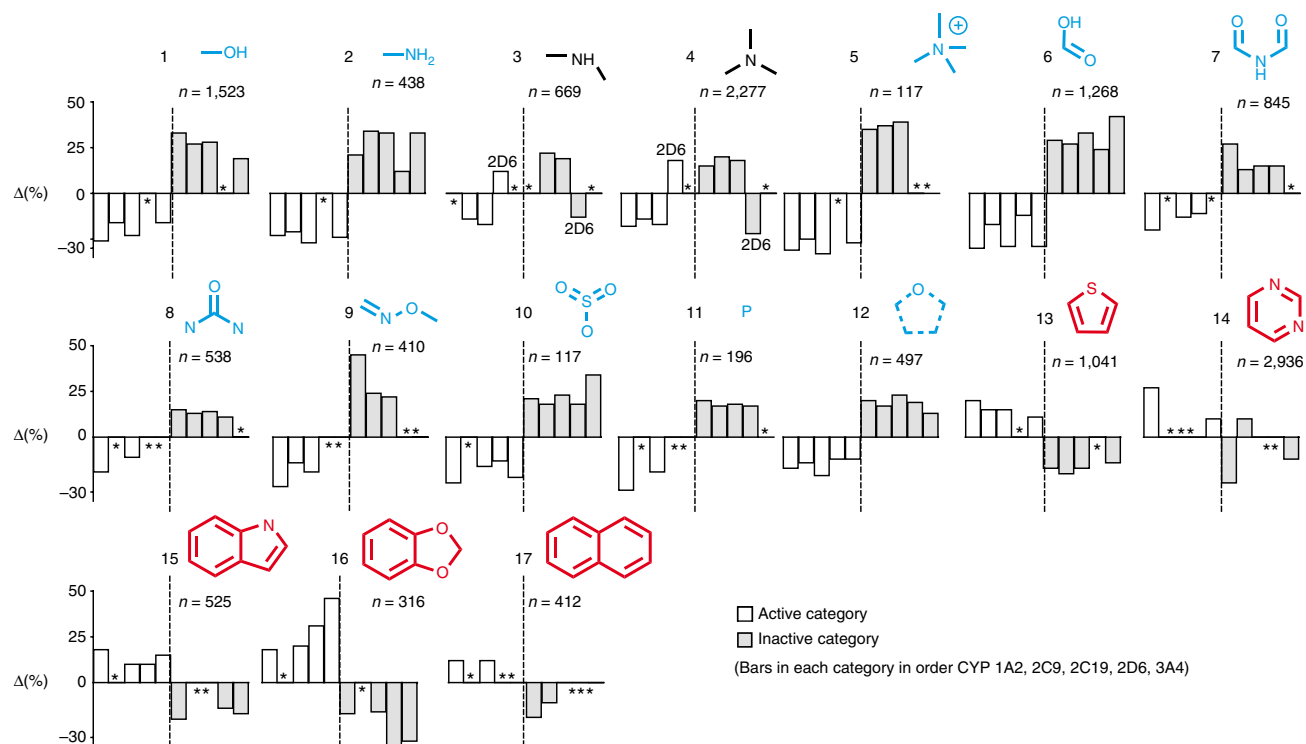


Figure 4 Fragment analysis of CYP activity. Shown are selected organic functional groups found to be disproportionately distributed among the response classes relative to the overall testing set. The entire testing set was 16,144 compounds with overall percentages of active and inactive compounds, respectively, of 36% and 42% (CYP 1A2), 25% and 50% (CYP 2C9), 36% and 42% (CYP 2C19), 15% and 66% (CYP 2D6), 32% and 44% (CYP 3A4). Data shown are the difference between the overall percentages and the distribution for the subset of compounds containing the indicated substructure ($\Delta\%$). A shift toward CYP inactivity is reflected in the graphs if the bars show negative percentage changes for the active class and positive for the inactive class. In contrast, a shift toward CYP activity is reflected by positive percentage changes in the active class and negative in the inactive class. Substructures are colored blue if these show shifts toward the pan-inactive class, and red if these show shifts toward the pan-active (inhibitor/substrate) class. Substructures showing isoform specificity (e.g., **3** and **4**) are colored black.

The presence of a carboxylic acid moiety (**6**) was also strongly correlated with pan-inactivating compounds. This can be compared with the trends for simple esters, amides and carbamates, which are generally much weaker and less consistent (data not shown). Imide (**7**) and urea (**8**) functionalities showed similar patterns to carboxylic acids, although with weaker associations. In the case of imides, the combinatorial library that contained a high density of this functionality (Fig. 5) may explain the weaker association. Oxime *o*-ethers (**9**), sulfonates (**10**) and phosphorus groups (**11**) were additional moieties associated with inactive compounds, although the frequency of occurrence of the latter in the testing set was low.

In terms of simple rings, the presence of oxolanes (**12**) was correlated with pan-inactive compounds. Aromatic equivalents such as thiophenes (**13**), furans and pyrroles (data not shown for the latter two) were associated with pan-activating compounds. Other aromatic groups such as pyrimidines, indoles, benzodioxoles and naphthalenes (**14** to **17**) were also generally associated with inhibitor/substrate compounds, particularly for CYP 1A2 (known to have preference for planar, polyaromatic substrates)⁶.

To identify more complicated substructures, we performed a similar analysis to that previously outlined¹⁰. For each isoform, inactive, high-confidence inhibitory and high-confidence activating compounds were clustered using extended-connectivity fingerprints (Pipeline Pilot 6.1, Scitegic, 2006, <http://accelrys.com/products/scitegic>). For each cluster the maximal common substructure was identified, which was then used to search the whole test set for similar structures. This process was repeated

several times with slightly different parameters for each isoform, in an attempt to ensure that a representative sample was taken.

The most prominent substructures from this analysis are summarized in Figure 5. We identified additional substructures associated with pan-inactive compounds, including long aliphatic carbon paths (**18**). Substructures incorporating simple chemical functionalities such as (**19**, **22**) showed isoform-specific behavior. Purine scaffold (**20**) and steroidal (**24**) compounds appeared to be largely inactive, consistent with the role of CYPs as largely metabolic rather than biosynthetic enzymes⁸. In contrast, a monosaccharide substructure (**23**) was associated with isoform-specific behavior—inactivity against CYP 1A2 but activity against CYP 3A4. We found that 89% of the compounds containing the quinazoline structure (**19**) were assigned to high-confidence inhibitory CRC for CYP 1A2, compared with only 8% for CYP 2C9. The known preferences of the CYP 1A and CYP 2C family of isozymes for planar, polyaromatic and nonplanar substrates, respectively⁶, explain this observation.

Scaffolds **21** and **22** were contained in two combinatorial libraries that were included in the testing set and shown in more detail in Supplementary Figure 5. The analysis method used here often required visual inspection to judge whether an identified substructure truly constituted an important common element. For example, the automated maximal common substructure procedure (and indeed the chemical group analysis) highlighted parts of the scaffolds shown in Supplementary Figure 5, but without visual inspection we would not have identified the corresponding combinatorial libraries.

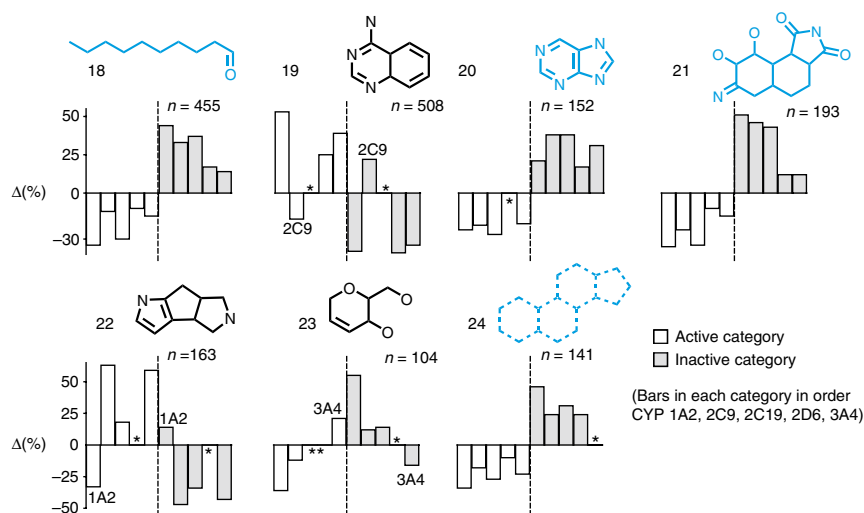


Figure 5 Fragment analysis of CYP activity for more complex heterocycles. Selected organic functional groups found to be disproportionately distributed among the response classes relative to the overall testing set. The data set is as described in **Figure 4** and the data shown are again the difference between the overall percentages and the distribution for the subset of compounds containing the indicated substructure ($\Delta\%$). Substructures are colored blue if these show shifts toward the pan-inactive class, and substructures showing isoform specificity are colored black. Entries where this value is less (in magnitude) than 10 are shown with a (*).

DISCUSSION

Using qHTS, we have defined pharmacological profiles of CYP activity with respect to libraries that included drugs, unoptimized commercially available compounds and combinatorial collections. A key advantage of this database is that the data are from a single series of experiments that used a bioluminescent assay format to determine potency for every compound. This allowed CRCs to be categorized to define activity and to directly compare results between isozymes. The CYP gene family has evolved to cover a wide range of chemical structures. We found that at least one isozyme was active in 93% of the MLSMR and 72% of the biofocused compounds. The database should aid in constructing and testing new predictive models of CYP activity.

We recognize that comparison of trends between isozymes provided here must be treated with care, as different probe substrates were used for the various isozymes (although all were based on luciferin), and this has previously been shown to influence observed effects on CYP activity^{22–24}. Overall, the bioluminescent assays demonstrated a correlation similar to that of CYP fluorescent assays when compared to conventional methods such as analytical detection of products (Online Methods). Excellent correlations were observed for CYP 1A2, 2C9 and 2D6 (r^2 values of 0.91, 0.84 and 0.94, respectively). CYP 2C19 and 3A4 also performed well, but were less well correlated (r^2 values of 0.54 and 0.74, respectively; **Supplementary Fig. 6**). Inhibitory activity in the assay may be due either to compounds acting as substrates or inhibitors, and some weak-binding substrates may be classified as ‘inactive’ (highest testing concentration was 57 μM). As our assay protocol did not include pre-incubation of compound with CYP, this database will be less sensitive toward time-dependent inhibitors and will miss mechanism-based inhibitors.

To assess if the potencies observed are clinically significant, we compared the IC_{50} to the C_{max} value (maximum plasma concentration of the drug) for ~140 drugs that inhibited one or more of the CYPs. From this analysis and based on FDA guidelines, we estimate that drug-drug interactions are probable for ~20% of the study

drugs showing inhibition, although the FDA criteria (drug-drug interactions are probable with $[\text{I}]/\text{K}_i > 0.1$) is stricter than what is typically applied in early optimization efforts (**Supplementary Fig. 7**).

Two isozymes, CYP 2C9 and 2C19, showed selectivity between drugs and the MLSMR. It has been suggested that CYP 3A4 is the most prominent P450 isozyme in drug metabolism and hepatic distribution (**Fig. 2b**)^{25,26}, but the drugs in our collection do not appear to have been optimized away from this activity. There has also been speculation that CYP 2D6 isozyme plays a prominent role in drug metabolism²⁷, but we did not observe differences in activity between MLSMR compounds and FDA-approved drugs for this isozyme. Our data show drugs to be more potent against CYP 2D6 than against the unoptimized compounds from the MLSMR, indicating that CYP 2D6 activity has not been an historical consideration in drug optimization efforts. Taken together, the CYP 2C family shows similar involvement in drug metabolism as CYP 3A4 (ref. 28), and CYP 2C9 shows a hepatic expression level similar to that of CYP 3A4 (**Fig. 2b**). Therefore, although activ-

ity against any or all CYP isozymes should be considered during lead optimization, the analysis provided here suggests that historically drugs have been particularly optimized against CYP 2C9 and 2C19 isozymes and that consideration of CYP 2C family activity should be particularly informative in drug optimization efforts.

Comparison of bulk compound properties between active and inactive compounds showed a slight differential for parameters related to aqueous solubility (**Supplementary Fig. 8**), consistent with trends shown in **Figures 4** and **5** such as the prevalence of aliphatic alcohol or charged groups among pan-inactives compounds. However, analysis of compound fragments also showed isozyme-selective substructures. Biochemicals such as steroids and purines were among the less active fragments, consistent with these five isozymes being primarily involved with xenobiotic metabolism. Common ring systems were more heterogeneous in their effects on CYP isozymes, with oxolanes showing a preference for pan-inactivity, whereas similar rings such as thiophenes or furans correlating with pan-activity.

Many computational strategies have been advanced to generate predictive CYP isozyme activity models^{29–33}. Several factors have been implicated in their limitations; foremost among these is the lack of a single large, diverse data set of CYP isozyme activities³⁴. We expect that the data set described here will facilitate the development of more-robust CYP activity models.

METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturebiotechnology/>.

Accession code. PubChem: assay identifier no. 1851.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This research was supported by the Molecular Libraries Initiative of the National Institutes of Health Roadmap for Medical Research and the Intramural Research

Program of the National Human Genome Research Institute. Work in Trinity College Dublin was supported by Enterprise Ireland, the Chemical Computing Group, OpenEye Scientific and Accelrys. We thank S. Jefferies and G. Carta for helpful discussions, S. Michael and C. Klumpp for help with robotic automation of the assays and P. Shinn for preparation of compound dilutions and library plates.

AUTHOR CONTRIBUTIONS

H.V. collected experimental data; H.V., N.S., R.H., T.J., D.F., N.A., M.S., D.G.L. and D.S.A. performed analysis; H.V., N.S., T.J., D.F., R.H., D.G.L., J.I., C.P.A. and D.S.A. wrote the paper.

Published online at <http://www.nature.com/naturebiotechnology/>.

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>.

- Evans, W.E. & Relling, M.V. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286**, 487–491 (1999).
- Ortiz de Montellano, P.R. (ed.) Cytochrome P450: structure, mechanism, and biochemistry. (Kluwer Academic/Plenum Publishers, New York, 2005).
- Williams, J.A. *et al.* Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUC/ AUC) ratios. *Drug Metab. Dispos.* **32**, 1201–1208 (2004).
- Di, L. & Kerns, E.H. Application of pharmaceutical profiling assays for optimization of drug-like properties. *Curr. Opin. Drug Discov. Devel.* **8**, 495–504 (2005).
- Hollenberg, P.F. Characteristics and common properties of inhibitors, inducers, and activators of CYP enzymes. *Drug Metab. Rev.* **34**, 17–35 (2002).
- Lewis, D.F., Eddershaw, P.J., Dickins, M., Tarbit, M.H. & Goldfarb, P.S. Structural determinants of cytochrome P450 substrate specificity, binding affinity and catalytic rate. *Chem. Biol. Interact.* **115**, 175–199 (1998).
- Porter, T.D. & Coon, M.J. Cytochrome P-450. Multiplicity of isoforms, substrates, and catalytic and regulatory mechanisms. *J. Biol. Chem.* **266**, 13469–13472 (1991).
- Sigel, A. *et al.* The Ubiquitous Roles of Cytochrome P450 Proteins: Metal Ions in Life Sciences vol. 3. (John Wiley & Sons, Ltd., 2007).
- Rendic, S. Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metab. Rev.* **34**, 83–448 (2002).
- Inglese, J. *et al.* Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci. USA* **103**, 11473–11478 (2006).
- Zheng, W. *et al.* Three classes of glucocerebrosidase inhibitors identified by quantitative high-throughput screening are chaperone leads for Gaucher disease. *Proc. Natl. Acad. Sci. USA* **104**, 13192–13197 (2007).
- Auld, D.S. *et al.* A basis for reduced chemical library inhibition of firefly luciferase obtained from directed evolution. *J. Med. Chem.* **52**, 1450–1458 (2009).
- Davis, R.E. *et al.* A cell-based assay for IκBα stabilization using a two-color dual luciferase-based sensor. *Assay Drug Dev. Technol.* **5**, 85–104 (2007).
- Xia, M. *et al.* Compound cytotoxicity profiling using quantitative high-throughput screening. *Environ. Health Perspect.* **116**, 284–291 (2008).
- Cali, J.J. *et al.* Luminogenic cytochrome P450 assays. *Expert Opin. Drug Metab. Toxicol.* **2**, 629–645 (2006).
- Lipinski, C.A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **44**, 235–249 (2000).
- Chauvet, N. *et al.* Description of a 96-well plate assay to measure cytochrome P4503A inhibition in human liver microsomes using a selective fluorescent probe. *Anal. Biochem.* **276**, 215–226 (1999).
- Kenworthy, K.E., Bloomer, J.C., Clarke, S.E. & Houston, J.B. CYP3A4 drug interactions: correlation of 10 *in vitro* probe substrates. *Br. J. Clin. Pharmacol.* **48**, 716–727 (1999).
- Kohonen, T. Self-organizing neural projections. *Neural Netw.* **19**, 723–733 (2006).
- Kohonen, T. & Oja, E. Computing with neural networks. *Science* **235**, 1227a (1987).
- Lewis, D.F.V. *A Guide to Cytochrome P450 Structure and Function* (Taylor & Francis, London, 2001).
- Foti, R.S. & Wahlstrom, J.L. CYP2C19 inhibition: the impact of substrate probe selection on *in vitro* inhibition profiles. *Drug Metab. Dispos.* **36**, 523–528 (2008).
- Kumar, V. *et al.* CYP2C9 inhibition: impact of probe selection and pharmacogenetics on *in vitro* inhibition profiles. *Drug Metab. Dispos.* **34**, 1966–1975 (2006).
- Nath, A. & Atkins, W.M. Principal component analysis of CYP2C9 and CYP3A4 probe substrate/inhibitor panels. *Drug Metab. Dispos.* **36**, 2151–2155 (2008).
- Shimada, T. *et al.* Cytochrome P450-dependent drug oxidation activities in liver microsomes of various animal species including rats, guinea pigs, dogs, monkeys, and humans. *Arch. Toxicol.* **71**, 401–408 (1997).
- Shimada, T., Yamazaki, H., Mimura, M., Inui, Y. & Guengerich, F.P. Inter-individual variations in human liver cytochrome P-450 enzymes involved in the oxidation of drugs, carcinogens and toxic chemicals: studies with liver microsomes of 30 Japanese and 30 Caucasians. *J. Pharmacol. Exp. Ther.* **270**, 414–423 (1994).
- Wienkers, L.C. & Heath, T.G. Predicting *in vivo* drug interactions from *in vitro* drug discovery data. *Nat. Rev. Drug Discov.* **4**, 825–833 (2005).
- Lewis, D.F. & Ito, Y. Human cytochromes P450 in the metabolism of drugs: new molecular models of enzyme-substrate interactions. *Expert Opin. Drug Metab. Toxicol.* **4**, 1181–1186 (2008).
- de Graaf, C., Pospisil, P., Pos, W., Folkers, G. & Vermeulen, N.P. Binding mode prediction of cytochrome p450 and thymidine kinase protein-ligand complexes by consideration of water and rescoring in automated docking. *J. Med. Chem.* **48**, 2308–2318 (2005).
- de Graaf, C., Vermeulen, N.P. & Feenstra, K.A. Cytochrome p450 *in silico*: an integrative modeling approach. *J. Med. Chem.* **48**, 2725–2755 (2005).
- Ekins, S., de Groot, M.J. & Jones, J.P. Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome p450 active sites. *Drug Metab. Dispos.* **29**, 936–944 (2001).
- Hansch, C., Leo, A., Mekapati, S.B. & Kurup, A. QSAR and ADME. *Bioorg. Med. Chem.* **12**, 3391–3400 (2004).
- Hansch, C., Mekapati, S.B., Kurup, A. & Verma, R.P. QSAR of cytochrome P450. *Drug Metab. Rev.* **36**, 105–156 (2004).
- Fox, T. & Kriegl, J.M. Linear quantitative structure-activity relationships for the interaction of small molecules with human cytochrome P450 isoenzymes. *Annual Reports in Computational Chemistry* vol. 3 (eds. Spellmeyer, D. and Wheeler, R.) 64–84, (Elsevier, New York, 2007).

ONLINE METHODS

Reagents. Luciferase-based P450-Glo Screening Systems were obtained from Promega for CYP 1A2 (V9770), CYP 2C9 (V9790), CYP2 C19 (V9880), CYP 2D6 (V9890), and CYP 3A4 Luciferin-PPXE (V9910) and were adapted for 1,536-well microplates and an automated protocol. Recombinant P450 enzymes were obtained from baculovirus constructs expressed in insect cells (BD/Gentest). The control compounds furafylline for 1A2, sulfaphenazole for CYP 2C9, ketoconazole for CYP 2C19, quinidine for 2D6, and ketoconazole for 3A4 were purchased from Sigma Aldrich. The 1,536-well microplates that were specifically bar coded for the Kalypsys automated robotic system were purchased from Greiner Bio-One.

Compound library and preparation of control plates. The 6,144 member biofocused library was collected from several sources and the composition of the compounds tested in this analysis were as follows: 31% National Cancer Institute, 18% LOPAC (Sigma Aldrich), 13% TOCRIS (TOCRIS), 18% Prestwick (Prestwick Chemical), 20% CMLD (Centers for Chemical Methodology and Library Development; 12% Boston University and 8% University of Pennsylvania). The biodiverse compound library consisted of 8,448 compounds (8,018 compounds are reported here) randomly selected from the larger 60,000 NIH MLSMR. Briefly, 10 mM stocks of compounds were diluted 7 times in a fivefold manner using an interplate dilution protocol; detailed library preparation is described elsewhere³⁵. Controls were added from a separate 1,536-well compound plate as follows: columns 1 and 2, 16-point titrations in duplicate of the appropriate inhibitor (both beginning at 10 mM in DMSO for all isozymes except CYP 2D6, for which the starting concentration was 250 μ M); column 3, the neutral control (DMSO); column 4, the appropriate control inhibitor (present at either 57 μ M or 1.4 μ M for CYP 2D6).

qHTS of CYP isozymes. The qHTS assay was performed in 1,536-well plates using an assay volume of 4 μ l, with a throughput of $\sim$ 7 samples/second. White 1,536-well solid bottom plates were incubated at the appropriate temperature (room temperature, $\sim$ 25 $^{\circ}$ C for CYP 1A2, 2D6, and 3A4 and at 37 $^{\circ}$ C for CYP 2C9 and 2C19). The Kalypsys dispenser was used to dispense 2 μ l enzyme and substrate mix. Then, control and compound collections were added with a Kalypsys pin tool through transfer of 23 nl of compound solution in DMSO. The plates were then incubated for 10 min at their respective incubation temperatures. The reaction was initiated through the addition of 2 μ l NADPH regeneration solution with the Kalypsys dispenser. The reaction was allowed to continue for 1 h before it was stopped with a detection reagent that contained luciferase and ATP to generate a luminescent signal. The plate was incubated for another 20 min before being read on a Viewlux (Perkin Elmer) with 2 $\times$ binning and an exposure of 60 s. The final 1,536-well assay protocol is summarized in **Supplementary Table 2**.

Classifying CRCs as inhibitory, activating or inactive. A total of 17,143 compounds had nonequivocal CRCs against all five isozymes, though more compounds were tested for each isozyme. Compounds with unavailable/equivocal data for even one CYP were not considered for analysis. This was done to have homogeneous data across the isozymes to facilitate comparisons between them. In-house software was used to classify the titration response by assigning curve classes based on shape, potency and fit of curve to the data. Class 1 constitutes a CRC that has lower and upper asymptote; class 2 has only an upper asymptote; class 3 shows activity only at highest concentration tested; class 4 shows no activity. As an additional criteria of activity we placed the CRCs into high-confidence (category 1) and low-confidence (category 2) CRCs based on the quality of the fit to the data^{10,36}. The category 1 CRC showed classes 1 or 2 CRCs with an efficacy of $\geq$ 60% inhibition. Any other activity was considered to have lower confidence and placed in category 2. Similarly, categories 3 and 4 designated high- and low-confidence activating CRCs.

Controlling for assay artifacts and aggregation-induced nonspecific inhibition. To address potential artifacts due to the assay format, particularly important for pan-active compounds, we used a database of potency values determined for the variant of the firefly luciferase used in the assay to remove any compounds that interfered with luciferase detection¹² (PubChem

AID: 1379; only 0.7% were found to be interfering in the compound collection described here). We also examined whether nonspecific inhibition due to compounds that form aggregates³⁷ occurred to a high degree in the CYP assays, a phenomenon that may be particularly relevant to the compounds that appear as activators of enzyme activity³⁸. However, we noted that CYP 3A4 activators didn't appear to be more or less selective or promiscuous than other compounds in our collection against multiple assays (data not shown). Also in line with this observation is that the activators observed for one CYP did not overlap with other CYPs supporting specific enzyme activation enzyme as opposed to a general nonspecific effect such as aggregation.

CYP enzymes are already packaged in membrane structures, which may help to mitigate aggregation-based nonspecific inhibition. For this reason, testing for detergent sensitivity could not be used as a counter-screen for colloidal aggregates due to greatly reduced activity of the CYP microsome preparations in the presence of detergent (both 0.1% and 0.01% TWEEN-20 and Triton were tried leading to a loss of luminescent signal).

As an alternative to detergents, to assess the effects of aggregation, we screened 44 randomly selected compounds against higher concentrations of CYP 3A4 and 2C9 (ref 39). Thirty-one compounds showed inhibition against CYP 3A4, and 29 showed inhibition against CYP 2C9. Compounds were titrated as 24-point twofold dilution series. We measured potencies in the bioluminescent assays under normal and high concentrations (100 nM, corresponding to 5 $\times$ and normal assay enzyme concentration for CYP 3A4 and 10 $\times$ for CYP 2C9). We then compared the IC₅₀s and Hill slopes of each compound at the two enzyme concentrations.

We observed an average shift of approximately two- to threefold (range 0.7- to fourfold; MSR for CYP 3A4 and 2C9 were 1.8 and 3.6, respectively). Only one compound showed a shift of fourfold in CYP 3A4 with two compounds showing a shift $\sim$ 4.5-fold in CYP 2C9. The average Hill coefficient was not greatly affected by the two conditions (values of 1.5/1.2 at normal enzyme concentration and 1.6/1.4 at high enzyme concentration, CYP 3A4/CYP 2C9, respectively). Therefore, irrelevant inhibitory effects due to compounds forming aggregates, which may bind and inhibit the enzyme with high affinity, do not appear to affect the CYP assays as conducted in this study, even though detergent was not present in the assay.

Testing of inhibitors and substrates. We examined several known substrates of CYP isozymes to determine the effect that these compounds have on pro-luciferin conversion (**Supplementary Fig. 2**). Phenacetin is the FDA-preferred substrate^{40,41} for *in vitro* CYP 1A2 studies, and theophylline and tacrine are acceptable FDA substrates for *in vitro* CYP 1A2 studies. Substrates such as phenacetin demonstrated AC₅₀s (that is, compound concentration that leads to 50% of the activity of an inhibition or activation control) that were within the range of reported K_Ms (e.g., AC₅₀ = 2.5 μ M, range 1.7–152 μ M; **Supplementary Fig. 2**). However, the highest tested concentration in the present assay was 57 μ M and substrates with weaker K_Ms showed only weak activity or were inactive in the assay, as in the example of theophylline (reported K_M values between 10 and 1,000 μ M; **Supplementary Fig. 2**).

Statistical assessment of qHTS CYP assay performance. A mean Z'-factor⁴² of 0.60 was determined for the biofocused library (Z'-factors for 1A2 = 0.69, 2C9 = 0.67, 2C19 = 0.35, 2D6 = 0.55, 3A4 = 0.72) whereas the MLSMR subset had a mean Z'-factor of 0.56 (Z'-factors for 1A2 = 0.74, 2C9 = 0.60, 2C19 = 0.16, 2D6 = 0.71, 3A4 = 0.60). For CRC-based screening the reproducibility of the AC₅₀s is the most relevant measure of data quality⁴³. For the present data set we calculated the minimum significant ratio (MSR)⁴⁴ for the control inhibitor titrations for all five isozymes. We found a mean MSR of 5.5 $\pm$ 4.4 with a mean s.d. of 0.23 $\pm$ 0.11 log AC₅₀ units (calculated using concentration of mol/l) in the biofocused collection and a mean MSR = 1.8 $\pm$ 0.4 with a mean s.d. of 0.08 $\pm$ 0.03 log AC₅₀ units for the MLSMR subset. To further validate the assays we obtained 91 randomly selected compounds for which bioassay data from another laboratory had also been published in PubChem (AIDs: 1024, 1025) and retested these in the CYP assays. A sample was considered 'confirmed' if it was found as active or inactive in both the original qHTS and confirmation experiments. We observed excellent confirmation (mean 86% $\pm$ 2; ranging between 84–90%) for all five isozymes (**Supplementary Fig. 1**).

Follow-up analysis. Follow-up compounds were ordered and prepared as 24-point titrations using a serial dilution (1:1) within the same plate. Plates were assayed using the same protocol as **Supplementary Table 1**, except that the plates were dispensed using a Multidrop (Thermo Scientific).

Analytical QC of compounds. The entire MLSMR library was subjected to purity analysis before plating (Discovery Partners International/Infinity Pharmaceutical). Active compounds that were obtained from commercial sources were reanalyzed for purity. For these resupplied compounds the purity analysis was performed via liquid chromatography–mass spectrometry (LC-MS) analysis on a Waters ACQUITY reversed-phase UPLC System and 1.7 μm bridged ethyl hybrid column (2.1 $\times$ 50 mm) using a linear gradient in 0.1% aqueous formic acid (5% acetonitrile in water increasing to 95% over 3 min). Compound purity was measured based upon peak integration from both UV/visible absorbance and evaporative light scattering detector readings, and compound identity was based upon mass analysis; all compounds passed purity criteria (>95%).

Processing of data to identify CYP selective fragments. The data were presented in a flat file format, where each record contained the tested substance and associated biological data. Where possible, the compound structure was recorded in SMILES format⁴⁵—records without this information, representing samples of unknown molecular composition such as ‘corn oil’, were eliminated from further consideration. Compound structures were placed in a canonical form with respect to charge/tautomer representation and protonation state. Common salts and counterions were stripped from the compounds, the protonation states were adjusted to those predicted at physiological pH using a substructure identification method (SciTegic 2006, Pipeline Pilot) and canonical tautomerization rules were applied (SciTegic 2006, Pipeline Pilot). At this point, records with identical structures, representing the same compound obtained from multiple sources, were merged. A set of simple organic functional groups were encoded as SMARTS strings (Daylight Theory Manual, Daylight Chemical Information Systems, <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>; cited 7/23/08) and used to search the set of test compounds. Where the same compound had been tested more than once on the same isoform the data were considered to be consistent if the same activity class was assigned in all cases, and, where applicable, the range of AC_{50} values did not exceed one log unit. In such cases, the average $\log\text{AC}_{50}$ was used as a representative value. Inconsistent data, as judged by these criteria, were discarded. The resulting data set contained 16,495 unique compounds. Of these, 16,144 were single-component structures, as determined by the absence of a period in the SMILES representation, containing only commonly occurring organic elements (H, C, N, O, P, S, Se, F, Cl, Br, I).

Self-organizing maps. Compound structures were converted into Daylight 2D fingerprints (Daylight Chemical Information Software, Daylight Chemical Information Systems), where each structure is encoded by 2,048 bits. Compounds were then clustered using the Self-Organizing Map (SOM) algorithm⁴⁶, with the software package downloaded from <http://www.cis.hut.fi/projects/somtoolbox/>, based on the similarity of their fingerprints (measured by the Euclidian distance) where each hexagon represents a cluster of structurally similar compounds, and compounds in neighboring hexagons (clusters) also share structural similarity. The hexagons in each SOM are colored by the enrichment level of active compounds (category 1 and 2) in each cluster

with respect to a CYP. Briefly, for each CYP, the fraction of active compounds in each cluster (structural class) was calculated and compared to the library average fraction of active compounds, and a statistical significance (P -value from a Fisher’s exact test where the null distribution is hypergeometric) calculated, where a smaller P -value indicates a higher significance level. The SOM is colored by the negative $\log_{10}$ of the P -values, such that a cluster with a dark red color is significantly enriched ($P < 0.05$), and a cluster with a dark blue color deficient, in active compounds when compared to the library average, and a green cluster either has no active compounds or its fraction of active compounds is not different from the library average.

Comparisons of CYP activity to conventional and C_{max} values. For comparison of the bioluminescent CYP data to conventional methods, the data were taken from references 22, 47, 48, which included human-liver microsomes assayed with isoform-specific probe substrates using high-performance LC or LC-MS methods of detection. For comparison of potency values to the C_{max} values were taken from the FDA web site or reference 49. The graphs shown in **Supplementary Fig. 7** are highlighted as described⁵⁰.

35. Yasgar, A. *et al.* Compound management for quantitative high-throughput screening. *J. Assoc. Lab. Autom.* **13**, 79–89 (2008).
36. Shukla, S.J. *et al.* Identification of pregnane X receptor ligands using time-resolved fluorescence resonance energy transfer and quantitative high-throughput screening. *Assay Drug Dev. Technol.* **7**, 143–169 (2009).
37. Feng, B.Y. *et al.* A high-throughput screen for aggregation-based inhibition in a large compound library. *J. Med. Chem.* **50**, 2385–2390 (2007).
38. Goode, D.R., Totten, R.K., Heeres, J.T. & Hergenrother, P.J. Identification of promiscuous small molecule activators in high-throughput enzyme activation screens. *J. Med. Chem.* **51**, 2346–2349 (2008).
39. Shoichet, B.K. Interpreting steep dose-response curves in early inhibitor discovery. *J. Med. Chem.* **49**, 7274–7277 (2006).
40. Eagling, V.A., Tjia, J.F. & Back, D.J. Differential selectivity of cytochrome P450 inhibitors against probe substrates in human and rat liver microsomes. *Br. J. Clin. Pharmacol.* **45**, 107–114 (1998).
41. von Moltke, L.L. *et al.* Phenacetin O-deethylation by human liver microsomes in vitro: inhibition by chemical probes, SSRI antidepressants, nefazodone and venlafaxine. *Psychopharmacology (Berl.)* **128**, 398–407 (1996).
42. Zhang, J.H., Chung, T.D. & 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).
43. Huang, R. *et al.* Characterization of diversity in toxicity mechanism using *in vitro* cytotoxicity assays in quantitative high throughput screening. *Chem. Res. Toxicol.* **21**, 659–667 (2008).
44. Eastwood, B.J. *et al.* The minimum significant ratio: a statistical parameter to characterize the reproducibility of potency estimates from concentration-response assays and estimation by replicate-experiment studies. *J. Biomol. Screen.* **11**, 253–261 (2006).
45. Weininger, D. Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
46. Kohonen, T. The self-organizing map. *Neurocomputing* **21**, 1–6 (1998).
47. Arimoto, R., Prasad, M.A. & Gifford, E.M. Development of CYP3A4 inhibition models: comparisons of machine-learning techniques and molecular descriptors. *J. Biomol. Screen.* **10**, 197–205 (2005).
48. Cohen, L.H., Remley, M.J., Raunig, D. & Vaz, A.D. *In vitro* drug interactions of cytochrome p450: an evaluation of fluorogenic to conventional substrates. *Drug Metab. Dispos.* **31**, 1005–1015 (2003).
49. Schulz, M. & Schmoldt, A. Therapeutic and toxic blood concentrations of more than 800 drugs and other xenobiotics. *Pharmazie* **58**, 447–474 (2003).
50. Zlokarnik, G., Grootenhuys, P.D. & Watson, J.B. High throughput P450 inhibition screens in early drug discovery. *Drug Discov. Today* **10**, 1443–1450 (2005).

QUERY FORM

NBT	
Manuscript ID	[Art. Id: 1581]
Author	
Editor	
Publisher	

AUTHOR:

The following queries have arisen during the editing of your manuscript. Please answer queries by making the requisite corrections directly on the galley proof. It is also imperative that you include a typewritten list of all corrections and comments, as handwritten corrections sometimes cannot be read or are easily missed. Please verify receipt of proofs via e-mail

Query No.	Nature of Query
Q1	Please carefully check the spellings of all author names and affiliations.
Q2	OK?
Q3	OK?
Q4	OK?
Q5	OK?
Q6	OK?
Q7	OK?
Q8	OK?
Q9	OK?
Q10	OK?
Q11	OK?
Q12	OK?
Q13	log10 OK? See query in SOM section of Online Methods
Q14	please verify that log10 was used because dark red on Fig 3 appears to correspond to $P=10^{-5}$, which is much smaller than 0.05. Current labels on color scale suggest that log2 was used. If log10, then shouldn't yellowish-orange correspond to $P=0.05$?
Q15	"OK?"