HHS Public Access

Author manuscript

Clin Pharmacol Ther. Author manuscript; available in PMC 2022 April 13.

Published in final edited form as:

Clin Pharmacol Ther. 2021 February ; 109(2): 485–493. doi:10.1002/cpt.2018.

Circulating miRNAs as biomarkers for CYP2B6 enzyme activity

Joseph Ipe[1,#], **Rudong Li**[2,#], **Ingrid F Metzger**[1,2], **Jessica Bo Li Lu**[1,2], **Brandon T Gufford**[1], **Zeruesenay Desta**[1,2], **Yunlong Liu**[2], **Todd C Skaar**[1,2]

[1]Division of Clinical Pharmacology, Indiana University School of Medicine, Indianapolis, IN 46202 USA

[2]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202 USA

Abstract

The *CYP2B6* gene is highly polymorphic and its activity shows wide inter-individual variability. However, substantial variability in CYP2B6 activity remains unexplained by the known *CYP2B6* genetic variations. Circulating, cell-free miRNAs may serve as biomarkers of hepatic enzyme activity. CYP2B6 activity in 72 healthy volunteers was determined using the disposition of efavirez as a probe drug. Circulating miRNA expression was quantified from baseline plasma samples. A linear model consisting of the effects of miRNA expression, genotype-determined metabolizer status, and demographic information was developed to predict CYP2B6 activity. Expression of 2510 miRNAs were quantified out of which 7 miRNAs, together with the CYP2B6-genotypic metabolizer status and demographics, was shown to be predictive markers for CYP2B6 activity. The reproducibility of the model was evaluated by cross-validation. The average Pearson's correlation (R) between the predicted and observed C_{max} ratios of efavirenz and its metabolite- 8-OH efavirenz using the linear model with all features (7 miRNA + metabolizer status + age + sex + race) was 0.6702. Similar results were also observed using AUC ratios (Pearson correlation's $R = 0.6035$). Thus, at least 36% (R^2) of the variability of *in vivo* CYP2B6 activity was explained using this model. This is a significant improvement over the models using only the genotype-based metabolizer status or the demographic information, which explained only 6% or less of the variability of *in vivo* CYP2B6 activity. Our results therefore demonstrate that circulating plasma miRNAs can be valuable biomarkers for *in vivo* CYP2B6 activity.

Keywords

Efavirenz; Metabolism; Pharmacokinetics; CYP

Corresponding Author: Todd C. Skaar, 950 W Walnut St. Room 419, Indianapolis, IN 46202.

[#]These authors contributed equally

Author Contributions:

J.I. and R.L. wrote the manuscript, J.I., Z.D., Y.L., and T.C.S. designed the research, J.I., R.L., I.F.M., and J.B.L.L. performed the research, J.I., R.L., B.T.G., and Y.L. analyzed the data.

Conflict of Interest: The authors declared no competing interests for this work.

Introduction:

Human CYP2B6 is primarily expressed in the liver and accounts on average for 2–5% of total hepatic CYP content. Over the past two decades, CYP2B6 has been shown to metabolize several xenobiotics and is estimated to metabolize about 8% of clinically used drugs such as bupropion, ketamine, meperidine, propofol, methadone, nevirapine, mephobarbital, cyclophosphamide, and efavirenz(1–4).

CYP2B6 expression has 20 to 250-fold inter- and intra-individual variability(5). Some of this variability can be attributed to genetic variations in the *CYP2B6* gene which is highly polymorphic with 38 alleles, several sub-alleles, and numerous sequence variants that have not been assigned a haplotype as of 2019(6, 7). CYP2B6*6 (516G>T and 785A>G) and CYP2B6*18 (983T>C, I328T) haplotypes result in a functionally deficient allele and include the most clinically relevant CYP2B6 polymorphisms identified, thus far (1, 8, 9). Variability in enzyme activity can also arise from a variety of additional factors that alter the transcriptional induction and suppression of the CYP2B6 gene. Induction of the CYP2B6 gene has been shown to be modulated through nuclear receptors, such as pregnane X receptor (PXR), constitutive androstane receptor (CAR), and glucocorticoid receptor (GR) (10–12). Thus, factors that activate or repress these receptors may have a significant effect on the downstream expression of CYP2B6 resulting in altered metabolism of its substrates.

Efavirenz, a non-nucleoside reverse transcriptase inhibitor used in the treatment of HIV type I, is a CYP2B6 substrate that is predominantly metabolized to 8-hydroxyefavirenz(13, 14). Efavirenz concentration in plasma has been shown to have extensive inter-individual variability. Subtherapeutic concentrations of efavirenz can lead to treatment failure while higher plasma concentrations observed in individuals with low-activity alleles have been associated with increased risk for central nervous system side effects(15, 16). The variability in enzyme activity is not fully attributable to the well-studied genetic polymorphisms in *CYP2B6*, suggesting that there may be additional factors such as miRNAs that contribute to the variability(4).

MicroRNAs are ~22 nucleotide non-coding RNAs that bind target sites predominantly in the 3’UTR of mRNAs and negatively regulate gene expression of majority of human genes including drug disposition genes. miRNAs are stably present in cell-free circulating form in biological fluids, such as plasma and serum that can be easily collected. There is ample evidence demonstrating the physiological and biological significance of miRNAs in regulating hepatic function through regulation of genes such as *HNF4A* (e.g. miR-449a), *PXR* (e.g. miR-30c-1–3p), and *CYP* enzymes(17–20). Studies profiling circulating miRNAs in liver diseases have shown that the expression of several miRNA measured in plasma are up or down regulated in liver diseases along with changes in CYP activity(18). *In vitro* studies support a role of miRNAs in the regulation of CYP2B6 (21). In addition, exchange of miRNA between liver and blood may result in an association between circulating miRNA concentrations and hepatic CYP2B6 activity.

Thus, we hypothesized that circulating plasma miRNAs are associated with hepatic CYP2B6 activity. The aim of this study was to identify miRNAs in human plasma that can predict

the CYP2B6 activity. To test our hypothesis, we used next generation sequencing to quantify plasma miRNAs in 72 healthy volunteers whose clinically relevant *CYP2B6* genotypes were determined. The CYP2B6 activity was measured using efavirenz as a probe drug and expressed as the $AUC_{0–48}$ (area under the curve plotted from 0–48 hours after dose) and C_{max} (highest plasma concentrations) ratios (8-hydroxy-efavirenz/efavirenz, 8-OH-EFV/EFV)(22). We adopted bioinformatic methods to develop a model to identify miRNAs that further explain CYP2B6 activity along with clinically relevant genotypes and demographic information. We show that circulating miRNAs are predictive for *in vivo* CYP2B6 activity.

Methods:

Measurement of CYP2B6 activity in healthy volunteers.

CYP2B6 enzyme activity was measured in 72 healthy volunteers, male (n=45) and female (n=27), ages 18–49, who were administered a single dose of the probe drug, efavirenz. The clinical trial was registered at ClinicalTrials.gov under identifier number NCT00668395. The study was approved by the Indiana University School of Medicine Institutional Review Board and the research was performed at Indiana University School of Medicine Clinical Research Center. Informed consent was obtained from the volunteers after which they underwent screening by review of medical history, physical examination, and laboratory tests such as electrocardiography, HIV test, urinalysis, and blood tests. Venous blood was collected for DNA isolation. Detailed inclusion/exclusion criteria and dietary restrictions have been previously reported(23, 24).

This study was open-label and designed to evaluate the pharmacokinetics, pharmacogenetics, and drug interactions after single and multiple doses of efavirenz. We used the data from the single dose arm of the study in which a 600 mg dose of efavirenz was administered orally. Plasma samples were taken at predose, 0.5, 1, 1.5, 2, 2.5, 3, 4, 6, 8, 10, 12, 16, 24, 48, 72 and 144h. Plasma concentrations of efavirenz and its metabolite 8-hydroxy-efavirenz was measured using liquid chromatography/tandem mass spectrometry (LC-MS/MS) as described in our previous publication(22). Time taken to reach maximum concentration (T_{max}) was of efavirenz and 8-OH-efavirenz was found to be ~3 hours. Pharmacokinetic parameters were estimated from plasma concentration-versus-time data by non-compartmental analysis, using Phoenix® WinNonlin® (version 7.0, Pharsight Corp., Cary, NC). Ratios of C_{max} (maximal plasma concentration) and $AUC_{0–48}$ (area under the plasma concentrations time curve from zero to 48h) of 8-hydroxyefavirenz to that of the parent drug served as markers of CYP2B6 activity.

C_{max} and AUC ratios showed the similar distribution and were highly correlated (Supplementary Figure S1). Thus, we used the C_{max} ratio as the primary phenotype of CYP2B6 activity in the context of modeling.

DNA genotyping

Genomic DNA extracted from pre-dose whole blood was used for genotyping. *CYP2B6* genotyping was performed using TaqMan Genotyping Assays for CYP2B6*6 [rs2279343, rs3745274] and CYP2B6*18 [rs28399499] (Life Technologies, Foster

City, CA) according to manufacturer's protocol, as previously described(23). *CYP2B6* rs2279343 (785A>G, K262R) is in a region that is identical to CYP2B7, a pseudogene. Genotyping of rs2279343 was performed by amplifying exon 5 with primers specific for CYP2B6 5′-CTCTCTCCCTGTGACCTGCTA-3′ (forward) and 5′-CTCCCTCTGTCTTTCATTCTGTC-3′ (reverse) (Integrated DNA Technologies, Coralville, IA) and using 1uL of the purified PCR amplification product as a template for a custom TaqMan Genotyping Assay, as previously described(23). Patient demographics and metabolizer status are listed in Supplementary Table S1. The observed minor allele frequencies were as expected. Metabolizer status based on genotype was determined as per Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline(25).

De-heparinization of RNA and library preparation

Venous blood was collected in tubes with heparin as the anti-coagulant. Plasma from such blood was aliquoted and stored at −80 °C. An aliquot was de-heparinized to avoid heparin mediated inhibition of polymerases used in next generation sequencing workflow. The plasma samples were thawed at room temperature and RNA was isolated using the miRNeasy Serum/Plasma kit according to the manufacturer's instructions. RNA (50 ng) was de-heparinized in a 20 μL reaction containing heparinase (6U), heparinase buffer, RNAse inhibitor (1 μL) [NEB, Ipswich, MA] and water. The reaction mixture was incubated at 30°C for 1 hour, followed by 99.9°C for 1 minute. The reaction was held at 4°C.

Sequencing library was prepared using QIAseq miRNA Library kit as per manufacturer's instructions. Sequencing was performed on Illumina-HiSeq with a read depth of 20M reads per sample.

miRNA quantification

Sequencing files were analyzed at Qiagen service labs using the GeneGlobe Data Analysis workflow (Qiagen, Frederick, MD). 3'-adapters were trimmed using 'cutadapt'(26) and reads with no adapter sequence are tallied. Following trimming, Unique Molecular Identifiers (UMIs) and insert sequences were identified. Reads with less than 16bp inserts and 10bp UMIs were discarded. To annotate the insert sequences, a unique sequence set was made for all readsets/samples in a submitted job. Following this, a sequential alignment strategy was followed to map to different databases (perfect match to miRBase mature, miRBase hairpin, noncoding RNA, mRNA and otherRNA, and ultimately a second mapping to miRBase mature, where up to two mismatches were tolerated) using 'bowtie'(27). miRBase V21 was used for miRNA, and piRNABank was used for piRNA.

Data normalization

Unique molecular identifier (UMI) read counts were used to quantify the expression of circulating miRNAs. However, the UMI read counts, like raw read counts in RNA-seq, have experimental sample effects. Thus, the UMI read counts were normalized against the RNA-seq library sizes. The normalization was performed using the algorithm of Trimmed Mean of M values (TMM), which is part of the R package "EdgeR" (v3.20.0) (28). In addition, no missing values (N.A.) of UMI count were observed for the sequenced miRNAs.

Linear regression model

We modeled the CYP2B6 activity by linear regression as the response to the expression of different miRNAs, clinically relevant genotypes (characterized as metabolizer status), and demographic information including age, sex and ethnicities. Each miRNA was regarded as an independent variable for CYP2B6 activity; other covariates, namely, metabolizer status, age, sex, and race were perceived as known (fixed) factors independent of the miRNAs. For example, if a set of miRNAs were associated with CYP2B6 activity (C_{max} ratio) along with other covariates, the formula of model was (Eqn.1):

$$C_{max}\left(\frac{8-OH-EFV}{EFV}\right) = \beta_0 + \beta_g[genotype\ ] + \beta_a[age\ ] + \beta_s[sex\ ] + \beta_r[race\ ] + \sum_k \beta_{miR}^{(k)} \cdot \left[miR^{(k)}\right] \quad \text{(Eqn. 1)}$$

where all independent variables were present in the formula additively as linear terms. 'C_{max}' was the quantified CYP2B6 activity (response variable); '*genotype*' was assigned the values of metabolizer status (poor, intermediate, and normal); each '$miR^{(k)}$' represented the (normalized) expression level of the k^{th} miRNA; the other covariates were '*age*', '*sex*' and '*race*'. Coefficients β_0, β_g, β_a, β_s β_r, and β_{miR} quantified the correlation/association of each of the regression variables to the response variable.

On the other hand, if we assumed that CYP2B6 activity was solely determined by the genotype-defined metabolizer status or demographic covariates, it would result in several other models: namely, the metabolizer status-only model (Eqn. 2), demographics-only model (Eqn. 3), or the miRNA-free model (Eqn. 4: combined model of Eqn. 2 and 3).

$$C_{max}\left(\frac{8-OH-EFV}{EFV}\right) = \beta_0 + \beta_g[genotype\ ] \quad \text{(Eqn. 2)}$$

$$C_{max}\left(\frac{8-OH-EFV}{EFV}\right) = \beta_0 + \beta_a[age] + \beta_s[sex] + \beta_r[race\ ] \quad \text{(Eqn. 3)}$$

$$C_{max}\left(\frac{8-OH-EFV}{EFV}\right) = \beta_0 + \beta_g[genotype\ ] + \beta_a[age\ ] + \beta_s[sex\ ] + \beta_r[race\ ] \quad \text{(Eqn. 4)}$$

Linear regression modeling was implemented in R (v3.4.4). The same approach was used when models were developed using AUC ratios as a measure of CYP2B6 activity.

Modeling

The workflow of the modeling approach is graphically shown in Figure 1. To establish a model, we preallocated a training and testing datasets respectively, and then implemented a supervised-feature selection (detailed as below).

Dataset partition: To construct the training and testing datasets, we randomly distributed the samples into the two categories. Specifically, 2/3 of the samples ($N_1 = 48$) were assigned

to the training set and the remaining 1/3 ($N_2 = 24$) were used as the testing data. We assigned the samples based on uniform probability-distribution, thus every sample had equal possibility of being distributed to either of the two sets. When building the model, the training set should be used to learn which features (i.e. miRNAs) could best explain the dataset, and thus the selected features and their respective coefficients were used as the (learned) model. Thereafter, this model was applied on the testing set to validate its performance on the new data not involved in the model-building. Distributions of features, as well as the C_{max} (and AUC) ratios across the training and testing datasets, were shown in Supplementary Figure S2.

To comprehensively test the model performance, we used the cross-validation method. In this respect, we resampled the training and testing datasets 200 times, i.e. randomly distributed a different 48-sample training set (L_i) and a 24-sample testing set (T_i) each time (i), and see how the selected features represented L_i and fitted T_i.

Supervision: In selecting features from the pool of a total of 2510 miRNAs, a supervision strategy was needed to shortlist the candidate features. This was to *i*) reduce the dimension of data input and thus alleviate the computing burden; and *ii*) enhance the algorithm convergence and thus accelerate the computation. Intuitively, we assumed that the top miRNAs that were individually explanatory to the enzyme activity were more informative as candidate features; and the number (N) of top miRNAs included was to be empirically determined in the supervision.

Using the training set (expression of 2510 miRNAs in 48 subjects), we first evaluated how each miRNA could explain the dataset, i.e. individually regressing $miRNA_k$ ($k = 1 \sim 2510$) to C_{max} ratios across the 48 subjects in the training set. Therefore, 2510 submodels and p-values were determined. Then the miRNAs were shortlisted with respect to two hyper-parameters, namely, the maximum number of features (i.e. N) and the cutoff for significance level of association (P) (29, 30). For example, if the actual number of significant miRNAs (individual p-value $\leq P$) was n; if $n \leq N$, we used the n miRNAs, and if $n > N$, we ranked these n miRNAs based on their degrees of associations (effect sizes), and the top N miRNAs were used for running the further feature selection (see below "Feature selection"). Empirically, we modulated P = e.g. 0.05, 0.01, 0.005, etc. and N = e.g. 10, 20, 30, etc. until the final resultant model achieved the highest stabilized performance on the training set (Figure 2A–2E, left panels). A stabilized performance implies that its level plateaus at a certain number of features. This criterion is required to rule out the possibility of overfitting, which implicates non-stable increase of model performance with respect to the number of features.

Models developed using the training dataset with different N and P values were then evaluated on the testing dataset (Figure 2A–2E, right panels). The other basal models with only the known features (e.g. metabolizer status, sex, age, or racial group) did not involve the supervision.

Feature selection: Training models were generated following feature selection from the pool of N candidate miRNAs shortlisted in the supervision step. We used Elastic Net (EN),

a regularized generalized linear regression algorithm, to select a subset of miRNAs that were non-redundant and continued to explain the training dataset optimally (31, 32). We utilized the R package CARET (Classification And REgression Training, version 6.0–79) to automate the process. CARET partitions the parameter space into a series of lattices and searches for the optimal regularization parameters according to changes of model fitness evaluated based upon a 3-fold cross validation implemented internally in the automatic process (33). Specifically, the optimal tuning parameters of EN here were $\alpha = 1$ (model fitness control) and $\lambda = 0.01023438$ (model flexibility control).

Noteworthy, dataset partition, supervision, or feature selection was indifferent with respect to the phenotypic activity metrics (C_{max} or AUC ratio) as the respective values were of the same numeric distribution. Based on all procedures above, we generated the optimally learned model that was comprised of the expression of 7 miRNAs, along with the known (non-miRNA) factors.

Model evaluation

Performance was scored as the *Pearson's Correlation Coefficient* (PCC) between the predicted values (i.e. model response) and experimentally observed values (i.e. measurement). Because we used a linear model, PCC^2 theoretically equals the percentage of model-explained variance in the total variance of real data. We applied the model on the training and testing sets independently to evaluate how well the model represented the data from which it was learned, as well as the performance on independent data. The PCCs from the training and testing sets were called the training and testing scores, respectively.

For the sake of generalizability, it is essential to evaluate the model robustness. A robust model is expected to have similar performance regardless of which parts of the sample set is used for training or testing; moreover, model coefficients calculated each time should be similar. Hence, we examined the respective performance levels (on each T_i, see the earlier "Dataset partition"), as well as the coefficient values calculated, based on each (different) L_i to see if they were similar or deviated in large. In addition, a well-learned model should exhibit balanced performances on the training and testing sets. Large difference between the training and testing scores indicates possible overfitting. In summary, we evaluated the model by checking the following aspects:

- **a.** performance of the randomly permuted cases was robust (i.e. falling in a specific range, which was quantified by standard deviation *sd* or standard error of mean *sem*).
- **b.** trends between training and testing scores were balanced.
- **c.** average levels of performance (on training and testing sets, respectively).
- **d.** feature coefficients calculated in the randomly permuted cases were similar/consistent (consistency quantified by standard deviation *sd*).

We applied this examination protocol to all the models in the study.

Results:

miRNA sequencing

A total of 2510 unique miRNAs were detected in the 72 samples. All 2510 miRNAs were included in the analysis. No predetermined UMI cutoff was used since the downstream supervision criteria will eliminate miRNAs expressed in limited samples and those that have low expression.

Prediction model for the CYP2B6 activity.

We have developed a model to predict CYP2B6 enzyme activity consisting of the expression of 7 circulating miRNAs and other covariates. The model is mathematically represented below (Eqn. 5).

$$C_{max}\left(\frac{8-OH-EFV}{EFV}\right) = \beta_0 + \beta_g[genotype\,] + \beta_a[age\,] + \beta_s[sex\,] + \beta_r[race\,] + \beta_1[miR_{204-5p}] + \beta_2[miR_{212-3p}] + \beta_3[miR_{3649}] + \beta_4[miR_{3941}] + \beta_5[miR_{4254}] + \beta_6[miR_{4442}] + \beta_7[miR_{6867-5p}] \quad \text{(Eqn. 5)}$$

Estimated values of the contributing coefficients β_0 (unknown factors), β_g (CYP2B6 genotype-characterized metabolizer status), $\beta_{a,\,s,\,r}$ (age, sex and race), and β_1 - β_7 (each selected miRNA) are listed in Supplementary Table S2. The miRNA effect sizes are specifically shown in Figure 3 with their respective 95% confidence intervals.

Robustness of these selected features was tested by resampling the training and testing datasets 200 times and examining the model performance for each time. The 7 miRNAs along with the other covariates were found to be highly robust as demonstrated by their presence as well as similar coefficient values observed over 200 randomizations. See Table S2 for the summary data of the randomizations for each miRNA as well as the other covariates. More specifically, R^2 (R-squared) values (for each instance) of the cross-validations were calculated and compared across the training and testing datasets (Supplementary Figure S3). Additionally, see Supplementary Table S3 for summary statistics of all the model features (as well as CYP2B6 activity) in the full dataset.

The training or testing score of the model is the Pearson's correlation coefficient (R) between the model-predicted and experimentally determined CYP2B6 activity on the training or testing datasets, respectively (Figure 4). The average training score (in 200 randomizations) was 0.7264 and the average testing score was 0.6702. Thus, the coefficient of determination for CYP2B6 activity by the model is 0.45 (R^2). In other words, 45% of the variability in CYP2B6 activity *in vivo* can be explained by our model.

To be comprehensive, data of modeling using the other CYP2B6 activity metric (AUC ratio) are shown in Supplementary Table S4.

Comparison of the full-feature model with alternative reduced models.

As a further means to prove the model validity and exclude the possibility of overfitting by excessive features, we computed the Akaike information criterion (AIC) values for all alternative models that are reduced by 1 feature, 2 features, etc. The values were shown in Supplementary Figure S4. Theoretically, $AIC = 2k - 2L(\hat{\theta})$, where k is the number of features and L is the value of maximum likelihood function for the fitted model $\hat{\theta}$). Generally, most alternative models of reduced features resulted in extra loss of information (i.e. elevated AIC).

Comparison of basal models with full-feature model.

The performance of our full-feature model was compared to the other models consisting of only the clinically genotypic CYP2B6 metabolizer status, demographic data (age, sex and ethnics) or both. Robustness of the metabolizer status-only and demographics-only models were also tested in 200 randomizations using the same protocol described above. Although, performance of the metabolizer status- and demographic-only models were found to be robust, the average training scores were only 0.2896 (metabolizer status-only), 0.2429 (demographic-only) and 0.3589 (metabolizer status + demographic), respectively; and the average testing scores were 0.2433, 0.1409 and 0.2182 respectively (Figure 4). Thus, only 6% of the variability in CYP2B6 activity could be explained by the metabolizer status-only model; and 2% could be explained by solely the demographic information. These results demonstrate that the use of expression levels of 7 circulating miRNAs significantly increases the predictive (or explanatory) power of CYP2B6 activity, *in vivo* (45%). One (average-level) example for the model accuracy, i.e. correlation between the predicted and experimental values of CYP2B6 activity, of the miRNA-featured model and the other miRNA-free models respectively (as validated results on the testing dataset), are shown in Figure 5.

Discussion:

CYP2B6 activity shows large inter-individual variability. While some of this variability can be explained by the known variants in the CYP2B6 gene, a large proportion of variability remains unaccounted for. In this study, we investigated the role of miRNAs as biomarkers of CYP2B6 function. There is constant exchange of miRNAs between blood and the different organs that it perfuses. Therefore, it is plausible that the *CYP2B6* expression in the liver may be regulated by miRNAs in circulation. It is also possible that the miRNAs secreted by the liver may be indicative of the level of CYP2B6 activity. We show that circulating miRNAs are associated with, and can be predictive of, *in vivo* CYP2B6 activity. Unlike a traditional approach aiming to identify associations between individual miRNAs and enzyme activity, our present study employed bioinformatic tools to build a ‘predictor model’ of CYP2B6 activity based on circulating miRNA levels. The model consists of a small subset of circulating miRNAs that robustly explained clinical CYP2B6 activity. Therefore, these miRNAs can be regarded as potential candidates for *in vivo* biomarkers.

In this study, the expression of 2510 miRNAs were quantified in 72 samples, out of which 7 miRNAs were shortlisted as biomarkers of CYP2B6 activity. The critical aspect of working

with such large amounts of data to identify biomarkers is the data interpretation because hundreds or thousands of associations will be discovered so we must find ways to prioritize these associations. Our bioinformatic modeling demonstrates that selection of candidate biomarkers through supervised feature selection can serve to prioritize the information that explains the phenotype. This practice can also be used on RNA sequencing datasets.

The algorithm identified seven circulating miRNAs that can predict *in vivo* CYP2B6 activity. The expression of each of these seven miRNAs is strongly correlated with CYP2B6 activity. However, these miRNAs together along with the clinically relevant genotypes have a more significant and reproducible predictive power than any individual miRNA or the metabolizer statuses alone. High correlations with CYP2B6 activity were observed for some of the individual miRNAs; but these correlations were lower than the combination of the 7 miRNAs and not as highly reproducible across randomization tests (Supplementary Figure S5). Therefore, the prediction model that includes the 7 miRNAs is not biased by any individual miRNA and the optimal prediction power is attained through the combination of the seven miRNAs. The seven miRNAs identified as predictors of CYP2B6 activity are miR-204–5p, miR-212–3p, miR-3649, miR-3941, miR-4254, miR-4442 and miR-6867–5p. Their respective levels of impact on CYP2B6 activity are summarized in Figure 3. Two of the seven miRNAs, miR-4254 and miR-6867–5p that are shown exerting positive and negative impact on the enzyme activity, respectively, have predicted binding sites on the CYP2B6 mRNA using TargetScan(34) and could plausibly regulate CYP2B6 activity through direct interaction with the transcript. The miRNAs may also function through regulating upstream regulators of CYP2B6 such as GR, PXR, and CAR. Two other miRNAs miR-204–5p (largest negative correlation with CYP2B6 activity) and miR-3941 (positive correlation) are predicted to target GR; and miR-4254 (positive correlation) is predicted also to target both PXR and CAR.

Our bioinformatic modeling aims to identify combination of independent features that are most explanatory to the observed phenotype. While there may be synergistic effects exerted by groups (or clusters) of miRNAs, each individual one may not significantly associate with phenotype, but may show explanatory power when clustered as a co-expression set. However, here we do not attempt to reveal the synergistic effect because this cannot be achieved by a *de novo* regularized feature-selection approach because there are too many possible combinations of miRNA clusters. Models like these could include known synergistic miRNA clusters should they be identified in the future.

Since the hyper-parameters of supervised modeling were empirically estimated, we implemented multiple tests to rule out model overfitting. To this respect, we have conducted: *i*) exploratory data analyses, which are prior examinations for features and datasets to ensure no bias were in the modeling data, *ii*) cross-validations, which are posterior tests for the robustness of predictive ability as well as model framework (i.e. feature and coefficients), and *iii*) systematic tests of information loss, which computed the AIC of our model against all possible alternative (reduced) models. On the other hand, the established model was also compared with all miRNA-free models (i.e. not featured by miRNA) to demonstrate the biological relevance of the selected miRNA features.

Nonetheless, one further issue is, although the model was validated based on the current sample set, the selected miRNAs need to be further investigated in functional experiments or at a larger population level for their validity as true clinically relevant biomarkers. Therefore, this present pilot research can serve as a foundation for future studies.

There is significant value in improved prediction of CYP2B6 activity in therapies using its substrates, such as efavirenz, which has shown to have adverse effects at high concentrations and treatment failure at too low concentrations. Although continued mechanistic investigation will be needed, we have demonstrated the value of circulating miRNAs as predictors of enzyme activity.A similar approach can be used for other CYP enzymes with unexplained variability.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: This study was supported by funding from NIH/NIGMS R-35 GM131812 (Skaar), Clinical Pharmacology Training Grant T32-GM008425 (Gufford), R01-GM078501, R01-GM121707 (Desta), and the Vera Bradley Foundation for Breast Cancer (Ipe)

References:

(1). Desta Z et al. Impact of CYP2B6 polymorphism on hepatic efavirenz metabolism in vitro. Pharmacogenomics 8, 547–58 (2007). [PubMed: 17559344]

(2). Eap CB et al. Stereoselective block of hERG channel by (S)-methadone and QT interval prolongation in CYP2B6 slow metabolizers. Clin Pharmacol Ther 81, 719–28 (2007). [PubMed: 17329992]

(3). Nakajima M et al. Genetic polymorphisms of CYP2B6 affect the pharmacokinetics/ pharmacodynamics of cyclophosphamide in Japanese cancer patients. Pharmacogenetics and Genomics 17, 431–45 (2007). [PubMed: 17502835]

(4). Zanger UM & Klein K Pharmacogenetics of cytochrome P450 2B6 (CYP2B6): advances on polymorphisms, mechanisms, and clinical relevance. Front Genet 4, 24 (2013). [PubMed: 23467454]

(5). Wang H & Tompkins LM CYP2B6: new insights into a historically overlooked cytochrome P450 isozyme. Curr Drug Metab 9, 598–610 (2008). [PubMed: 18781911]

(6). Gaedigk A et al. The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. Clin Pharmacol Ther 103, 399–401 (2018). [PubMed: 29134625]

(7). Gaedigk A et al. The Evolution of PharmVar. Clin Pharmacol Ther 105, 29–32 (2019). [PubMed: 30536702]

(8). Klein K et al. Genetic variability of CYP2B6 in populations of African and Asian origin: allele frequencies, novel functional variants, and possible implications for anti-HIV therapy with efavirenz. Pharmacogenetics and Genomics 15, 861–73 (2005). [PubMed: 16272958]

(9). Hofmann MH et al. Aberrant splicing caused by single nucleotide polymorphism c.516G > T Q172H, a marker of CYP2B6*6, is responsible for decreased expression and activity of CYP2B6 in liver. Journal of Pharmacology and Experimental Therapeutics 325, 284–92 (2008).

(10). Wang H et al. A novel distal enhancer module regulated by pregnane X receptor/constitutive androstane receptor is essential for the maximal induction of CYP2B6 gene expression. J Biol Chem 278, 14146–52 (2003). [PubMed: 12571232]

(11). Wang H et al. Glucocorticoid receptor enhancement of pregnane X receptor-mediated CYP2B6 regulation in primary human hepatocytes. Drug Metab Dispos 31, 620–30 (2003). [PubMed: 12695351]

(12). Sueyoshi T, Kawamoto T, Zelko I, Honkakoski P & Negishi M The repressed nuclear receptor CAR responds to phenobarbital in activating the human CYP2B6 gene. J Biol Chem 274, 6043–6 (1999). [PubMed: 10037683]

(13). Ward BA, Gorski JC, Jones DR, Hall SD, Flockhart DA & Desta Z The cytochrome P450 2B6 (CYP2B6) is the main catalyst of efavirenz primary and secondary metabolism: implication for HIV/AIDS therapy and utility of efavirenz as a substrate marker of CYP2B6 catalytic activity. J Pharmacol Exp Ther 306, 287–300 (2003). [PubMed: 12676886]

(14). In: Consolidated Guidelines on the Use of Antiretroviral Drugs for Treating and Preventing HIV Infection: Recommendations for a Public Health Approach (ed. nd) (Geneva, 2016).

(15). Adkins JC & Noble S Efavirenz. Drugs 56, 1055–64 (1998). [PubMed: 9878993]

(16). Marzolini C, Telenti A, Decosterd LA, Greub G, Biollaz J & Buclin T Efavirenz plasma levels can predict treatment failure and central nervous system side effects in HIV-1-infected patients. Aids 15, 71–5 (2001). [PubMed: 11192870]

(17). Nakajima M & Yokoi T MicroRNAs from biology to future pharmacotherapy: Regulation of cytochrome P450s and nuclear receptors. Pharmacol Therapeut 131, 330–7 (2011).

(18). Vuppalanchi R et al. Relationship between differential hepatic microRNA expression and decreased hepatic cytochrome P450 3A activity in cirrhosis. Plos One 8, e74471 (2013).

(19). Ramamoorthy A et al. In Silico and In Vitro Identification of MicroRNAs That Regulate Hepatic Nuclear Factor 4 alpha Expression. Drug Metabolism and Disposition 40, 726–33 (2012). [PubMed: 22232426]

(20). Vachirayonstien T & Yan B MicroRNA-30c-1–3p is a silencer of the pregnane X receptor by targeting the 3'-untranslated region and alters the expression of its target gene cytochrome P450 3A4. Biochim Biophys Acta 1859, 1238–44 (2016). [PubMed: 27085140]

(21). Burgess KS et al. Variants in the CYP2B6 3′ UTR Alter In Vitro and In Vivo CYP2B6 Activity: Potential Role of MicroRNAs. Clinical Pharmacology & Therapeutics 104, 130–8 (2018). [PubMed: 28960269]

(22). Robarge JD et al. Population Pharmacokinetic Modeling To Estimate the Contributions of Genetic and Nongenetic Factors to Efavirenz Disposition. Antimicrobial Agents and Chemotherapy 61, o (2017).

(23). Michaud V et al. Efavirenz-mediated induction of omeprazole metabolism is CYP2C19 genotype dependent. Pharmacogenomics Journal 14, 151–9 (2014).

(24). Michaud V et al. Induction of CYP2C19 and CYP3A Activity Following Repeated Administration of Efavirenz in Healthy Volunteers. Clin Pharmacol Ther 91, 475–82 (2012). [PubMed: 22318618]

(25). Desta Z et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for CYP2B6 and Efavirenz-Containing Antiretroviral Therapy. Clin Pharmacol Ther 106, 726–33 (2019). [PubMed: 31006110]

(26). Martin M Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* 17, 3 (2011).

(27). Langmead B, Trapnell C, Pop M& Salzberg SL Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009). [PubMed: 19261174]

(28). Robinson MD & Oshlack A A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11, R25 (2010). [PubMed: 20196867]

(29). Tibshirani R, Hastie T, Narasimhan B & Chu G Diagnosis of multiple cancer types by shrunken centroids of gene expression. P Natl Acad Sci USA 99, 6567–72 (2002).

(30). Hastie T, Tibshirani R, Narasimhan B & Chu G Supervised learning from microarray data. Compstat 2002: Proceedings in Computational Statistics, 67-77 (2002).

(31). Friedman J, Hastie T & Tibshirani R Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33, 1–22 (2010). [PubMed: 20808728]

(32). Zou H & Hastie T Regularization and variable selection via the elastic net. J R Stat Soc B 67, 301–20 (2005).

(33). Kuhn M Building Predictive Models in R Using the caret Package. Journal of Statistical Software 28, 1–26 (2008). [PubMed: 27774042]

(34). Agarwal V, Bell GW, Nam JW & Bartel DP Predicting effective microRNA target sites in mammalian mRNAs. Elife 4, (2015).

Study Highlights:

What is the current knowledge on the topic?

CYP2B6 is a highly polymorphic enzyme that has a wide pharmacokinetic variability. However, all of this variability cannot be accounted for by demographic information and the known allelic variants *CYP2B6*6* and *CYP2B6*18*. Circulating miRNAs are constantly exchanged between blood and different organs in the body including the liver.

What question did this study address?

We hypothesized that circulating miRNAs are associated with hepatic CYP2B6 activity. We measured baseline miRNA expression in plasma from 72 healthy volunteers whose CYP2B6 enzyme activity was measured using the probe drug efavirenz. We determined whether the expression of certain miRNAs was associated and predictive of CYP2B6 activity *in vivo.*

What does this study add to our knowledge?

We developed a model consisting of seven miRNAs along with demographic information and the CYP2B6 metabolizer status that could predict CYP2B6 enzyme activity, *in vivo*. The model with seven miRNAs, demographics and metabolizer status was able to explain 36% of the variability in CYP2B6 activity while models without miRNAs were able to explain only 6% or less.

How might this change clinical pharmacology or translational science?

The identification of seven miRNAs that are associated with CYP2B6 activity helps further explain the pharmacokinetic variability in CYP2B6 in addition to the known variant genotypes. Measuring the expression of these miRNAs along with the genotyping patients may help improve the prediction of CYP2B6 activity. This study also demonstrates a novel modelling approach to correlate miRNA expression to CYP2B6 enzyme function. This approach can be used for other enzymes with unexplained pharmacokinetic variability.

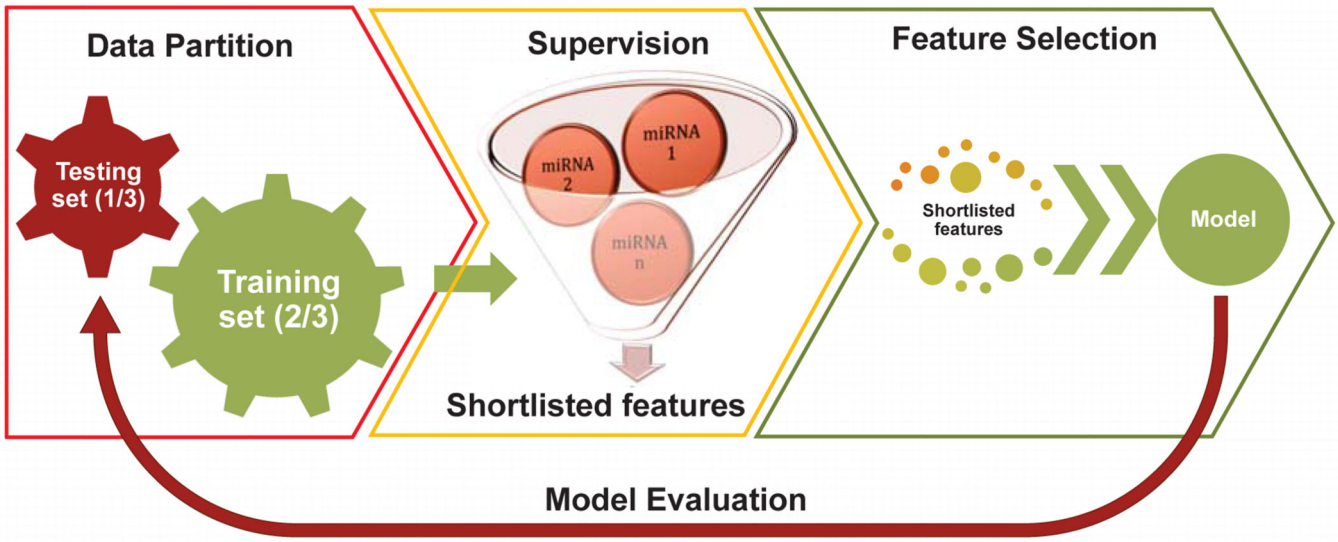

Figure 1.
Workflow of modeling showing data partition, supervision and feature selection followed by model evaluation.

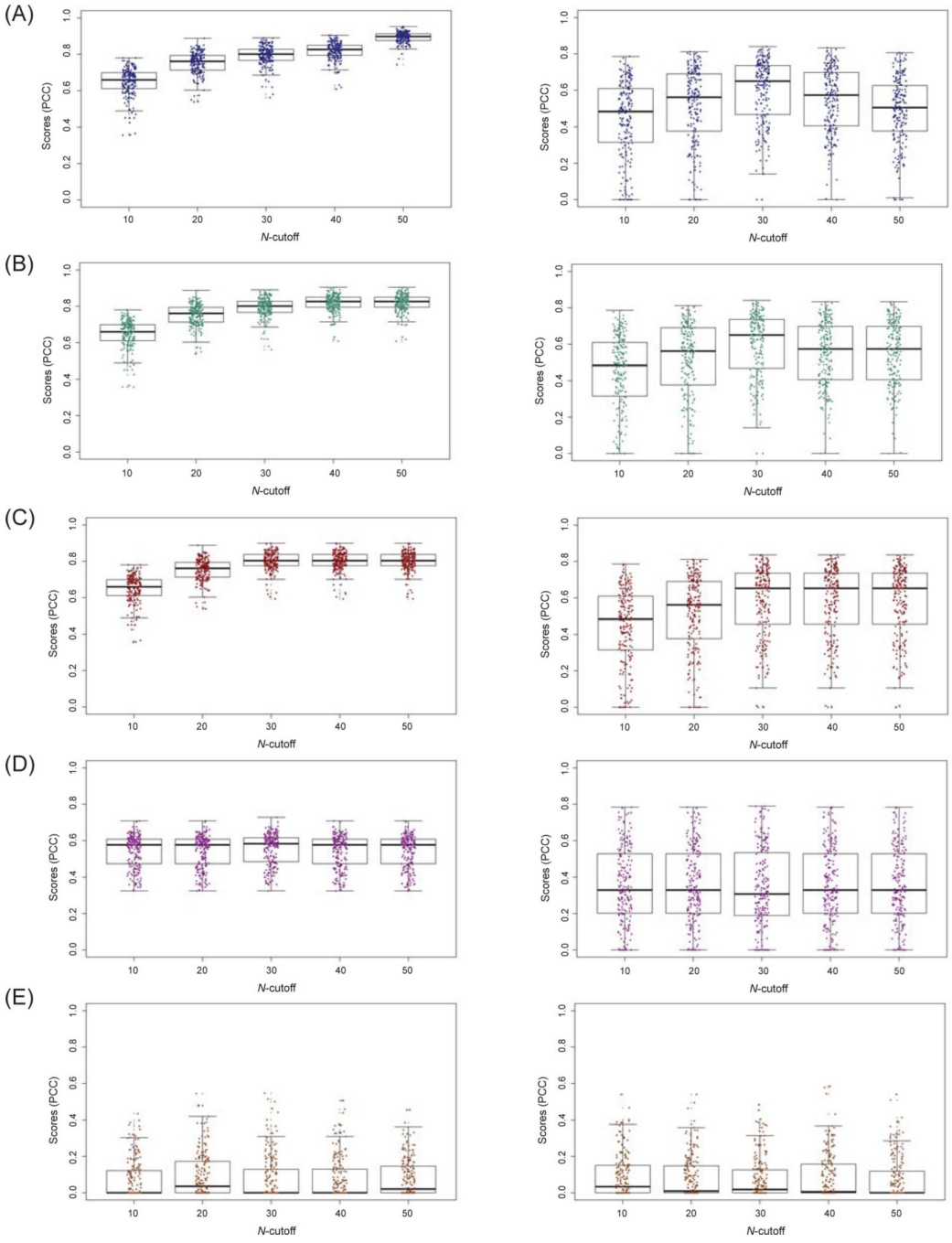

Figure 2: Results of numerical experiments on supervision cutoffs

Values in the Y-axis are the performance scores (Pearson's correlation) of the models. X-axis denotes the maximum number of features (*N*) included in the model. A- E denotes p-value cutoffs (*P*) of 0.05, 0.01, 0.005, 0.001, and 0.0005. The plots on the left show the performance of the training model, and the model evaluation on unused data is shown on the right.

The *P* of 0.005 and *N* of 30 resulted in a learning curve with the highest stable score on the training set; and showed optimal performance on the testing set.

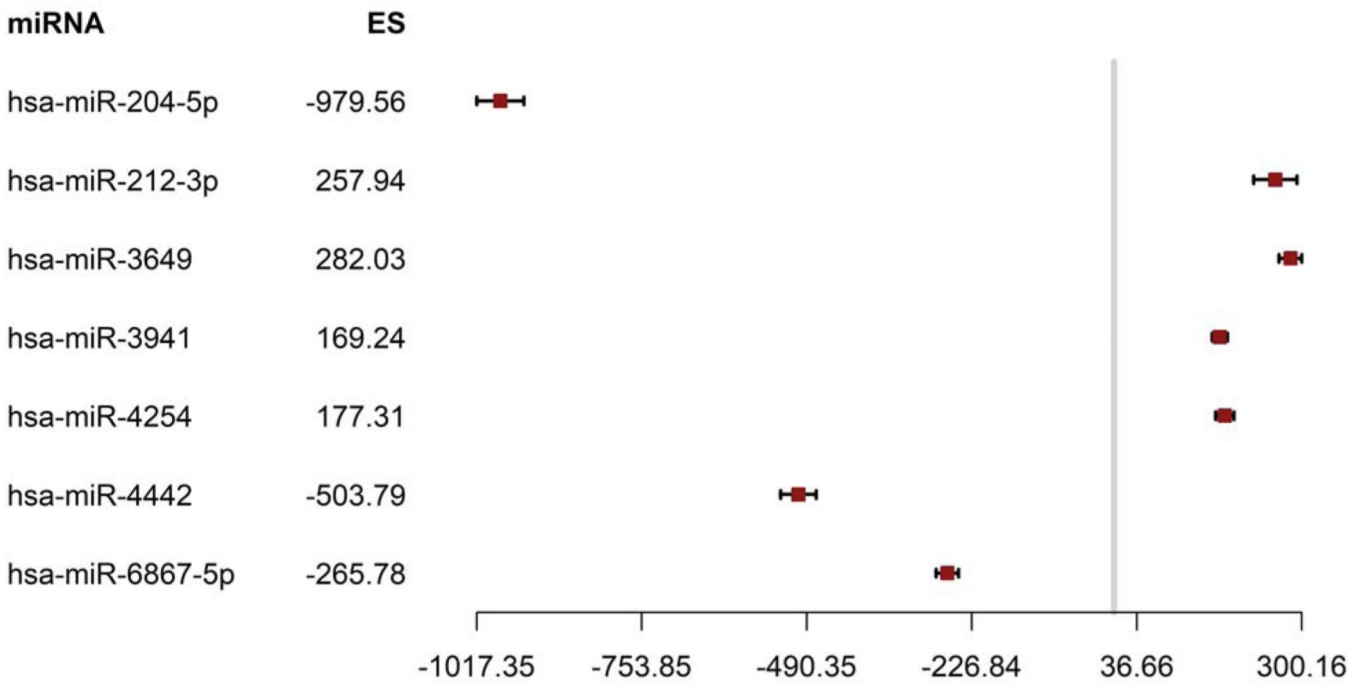

Figure 3. Forest plot of the selected miRNAs with effect sizes on CYP2B6 activity.
Effect sizes (estimated) of the 7 miRNAs included in the model are shown with 95% confidence intervals.

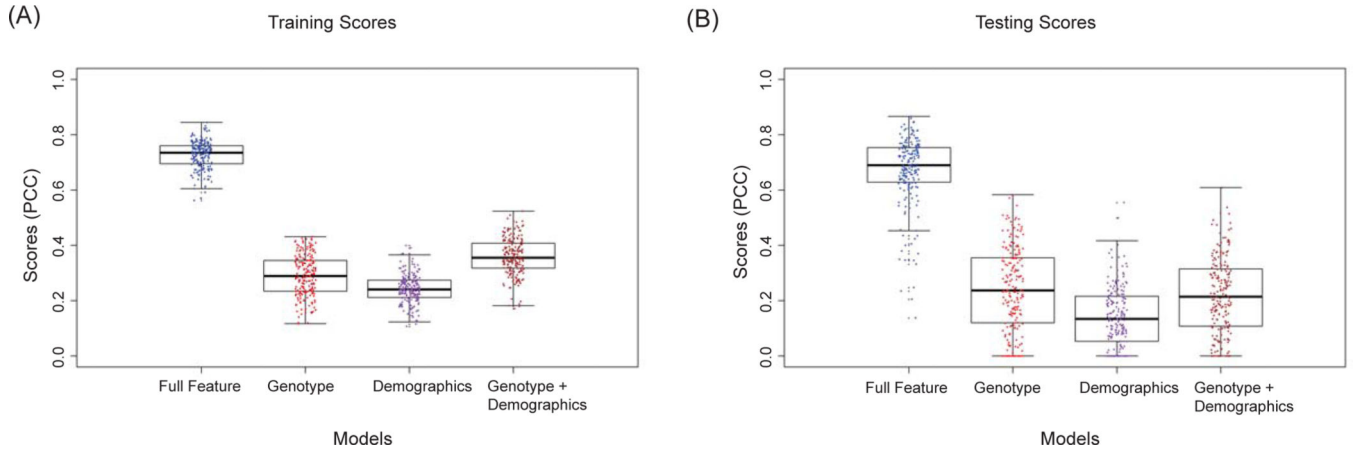

Figure 4. Performance of prediction model.

Performance is scored as the Pearson's correlation between predicted and observed values, the results of 200 randomized tests on the full feature model (7 miRNAs + metabolizer status + age + sex + race), metabolizer status-only, demographic only, and both metabolizer status and demographic are shown.

(A): Training score -performance of the full feature model and other models on the training set. Average level for full feature model = 0.73 (std. error of mean = 0.004), metabolizer status-only model = 0.29 (std. error of mean = 0.005), demographics only model = 0.24 (std. error of mean = 0.004), and the metabolizer status + demographics model = 0.36 (std. error of mean = 0.005).

(B): Testing score -performance of the full feature model and other models on the testing set. Average level for full feature model = 0.67 (std. error of mean = 0.009), metabolizer status-only model = 0.24 (std. error of mean = 0.011), demographics only model = 0.13 (std. error of mean = 0.009), and the metabolizer status + demographics model = 0.21 (std. error of mean = 0.01). Bold line indicates the median, the box spans the lower (1^{st}) and upper (3^{rd}) quartile, and the whisker bars extent to the minimum and maximum values. Data points marked by "+" extending beyond the whiskers were the outliers.

As shown, (average) training and testing scores are similar, indicating no or negligible overfitting. Additionally, most of the scores fall in the specific range marked by the boxplots, suggesting robust model performance. The small standard errors indicate accurate estimation of the means, i.e. most scores stably reside in a narrow range.

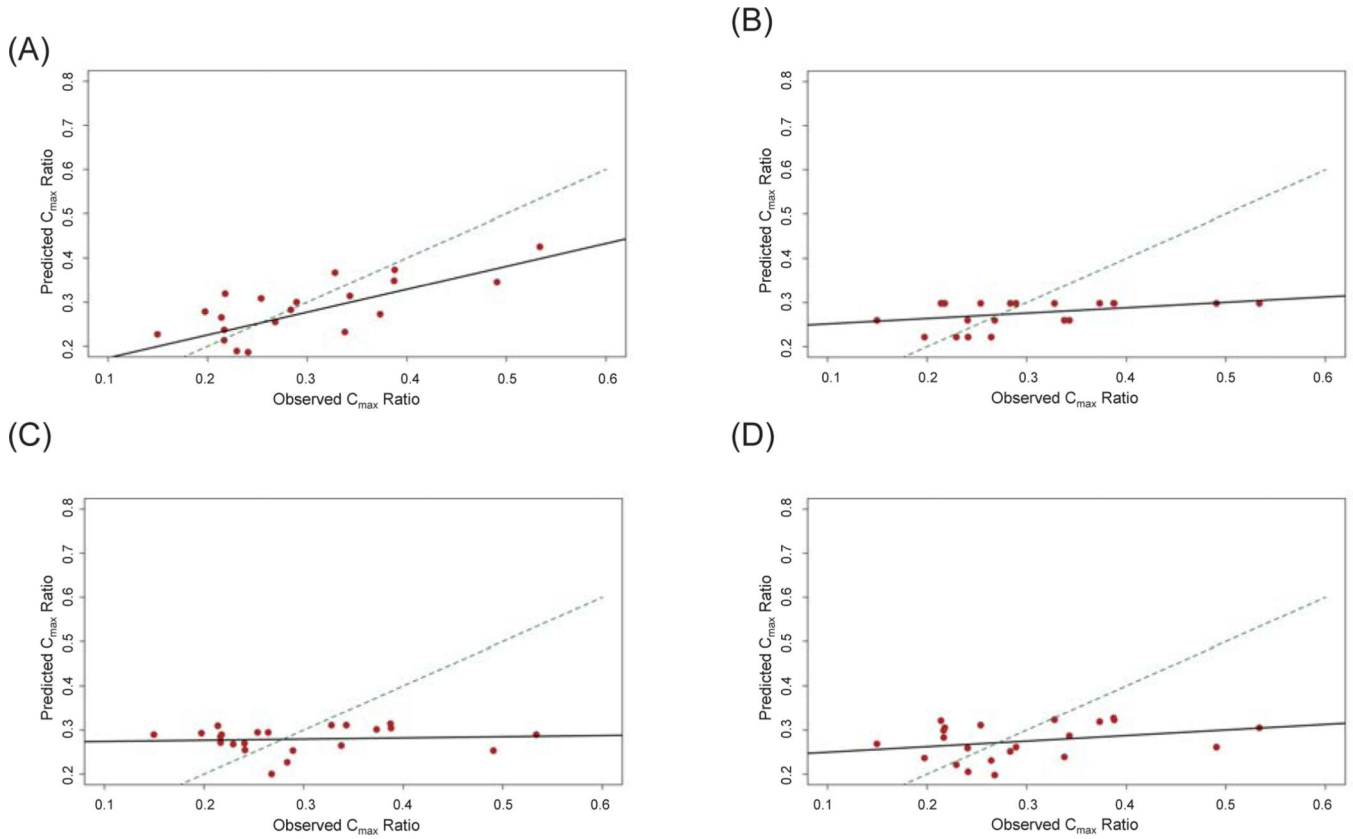

Figure 5. Model performance showing correlation between predicted and experimental values.

One example (from 200 randomizations) of correlation between CYP2B6 activity predicted by the models and experimental values in the testing data, using an instance from the 200 randomizations that was closest to the average performance. A) Full feature model, B) Metabolizer status only, C) Demographics only, and D) Metabolizer status + Demographics model. Dashed line is the line of identity and the solid black line is the regression line.