Hubble2D6: A deep learning approach for predicting drug metabolic activity

Gregory McInnes[1], Rachel Dalton[2,3], Katrin Sangkuhl[4], Michelle Whirl-Carrillo[4], Seung-been Lee[5], Russ B. Altman[4.6]*, Erica L. Woodahl[2]*

[1]Biomedical Informatics Training Program, Stanford University, Stanford, CA

[2]Department of Biomedical and Pharmaceutical Sciences, University of Montana, Missoula, MT

[3]Department of Biomedical and Translational Research, University of Florida, Gainesville, FL

[4]Department of Biomedical Data Science, Stanford University, Stanford, CA

[5]Department of Genome Sciences, University of Washington, WA

[6]Departments of Genetics, and Medicine, Stanford University, Stanford, CA

*Corresponding author. Email: rbaltman@stanford.edu (R.B.A.), erica.woodahl@umontana.edu (E.L.W.)

Abstract

A major limitation of phenotype prediction in genetics is the ability to model the complexities of genetic variation when sample sizes are small. This is especially true in pharmacogenetics, a highly translational yet data-limited subfield of genetics. Drug metabolism is a critical facet of pharmacogenetics and can have consequences for drug safety and efficacy. *CYP2D6* is an important enzyme, metabolizing more than 25% of clinically used drugs. It is highly polymorphic which leads to a heterogeneous response to drugs among the population. We present Hubble2D6, a set of deep learning models for predicting metabolic activity of *CYP2D6* genotype and predicting functional classification of *CYP2D6* haplotypes. We train our models on 249 samples, addressing data scarcity by pretraining on simulated data, weakly supervised learning, and using a functional representation of genetic variants. We validate our models using *in vitro* data for haplotypes previously unseen by the model and explain 38% of the variance with the genotype-based activity predictor and predict haplotype function with an AUC of 0.85. We demonstrate a procedure to build a computational model of a complex gene using primarily simulated and unlabeled data which can then be used to make functional predictions about novel genetic variation, and present a model that may be of clinical significance for an important application of genetics.

Introduction

Pharmacogenomics has emerged as one of the most clinically actionable subfields of modern genetics. With proper use, pharmacogenomics can offer clinical guidance to clinicians to provide personalized drug selection and dosing to patients based on genomic markers that predict how they will respond to drugs. The Clinical Pharmacogenomics Implementation Consortium (CPIC) has issued clinical guidelines for 47 drugs. These guidelines may suggest an alternate dose or a different drug altogether based on an individual's genetics[1,2]. Studies have shown that as many as 99% of individuals carry at least one actionable pharmacogenetic variant that could lead to prescribing changes in at least 1 medication[3–5].

Cytochrome P450 family 2, subfamily D, polypeptide 6 (*CYP2D6),* is one of the most important pharmacogenes. The protein, a hepatic enzyme dimer localized in the endoplasmic reticulum, metabolizes more than 25% of clinically used drugs including antidepressants, antipsychotics, opioids, antiemetics, antiarrhythmics, β-blockers, and cancer chemotherapeutics[6,7]. In Nordic European countries as many as 16% of inhabitants are on at least one drug metabolized by CYP2D6[5].

Adding to its clinical importance, *CYP2D6* is highly polymorphic[8]. More than 130 haplotypes (known as star alleles) comprised of single nucleotide variants (SNVs), insertions and deletions (INDELs), and structural variants (SVs) have been discovered and catalogued in Pharmacogene Variation Consortium (PharmVar; www.pharmvar.org), many of which are known to alter functional activity[9,10]. Levels of enzymatic activity and gene expression are influenced by changes to the DNA in the *CYP2D6* locus. Individuals are typically broken into one of four metabolizer classes that define *CYP2D6* metabolic phenotypes: normal (NM), intermediate (IM), poor (PM), and ultrarapid metabolizers (UM). Frequency of metabolizer classes varies widely among global populations; PM range from 0.5% to 5.4%, IM range from 2.8% to 11%, and UM range from 1.4% to 21.2%[11].

CYP2D6 is also prone to structural variation, which can both increase and decrease function[12]. Copy number variants and hybridizations with the pseudogene *CYP2D7* are observed. The Estonian Biobank study found that 4.1% of the participants had a copy number aberration[5]. In addition to *CYP2D6*, the locus contains two non-functional pseudogenes, *CYP2D7* and *CYP2D8*. *CYP2D6* and *CYP2D7* have a homologous repetitive region downstream which leads to gene hybridizations as a result of unequal recombination. Approximately 2% of individuals harbor a *CYP2D6-2D7* hybrid tandem arrangement which negatively impacts enzymatic function[13]. Structural variation in *CYP2D6* has strong effects on phenotype: gene duplications of functional haplotypes can lead to an individual being an UM, and *CYP2D6-2D7* hybridizations lead to a non-functional copy of the gene which decreases overall metabolic activity. Structural variants have been shown to explain 5% of the variance in metabolic activity of *CYP2D6*[14].

Despite its highly polymorphic nature, *CYP2D6* is one of the most clinically actionable pharmacogenes. Clinical guidelines providing dosage recommendations for different metabolizer classes have been published by CPIC for drugs that are metabolized by CYP2D6. Among the drugs with guidances are ones that are frequently prescribed, including opioids and antidepressants. Following the guidelines for these drugs could improve patient outcomes by decreasing adverse effects or increasing efficacy. It has been suggested that using pharmacogenetics guided opioid therapy could be part of a solution for combating the opioid epidemic[15].

Pharmacogenetic dosing guidances presume that the clinician has access to the patient's *CYP2D6* genotype and that the resulting phenotype can be correctly predicted. There are previous efforts to predict CYP2D6 function of haplotypes, the best known is the activity score system (AS)[16]. The AS works by assigning a value to each haplotype (0 for no function alleles, 0.5 for decreased function, and 1 for normal function) then summing the scores of all *CYP2D6* star alleles observed in an individual's genome. The resulting AS for the person's genotype is used to determine the CYP2D6 phenotype. AS value assignment for the haplotypes relies heavily on the manual curation of known star alleles through a review of the literature. Most often, *in vitro* experiments and *in vivo* clinical outcomes are used to make a determination of star allele function. Several tools exist to determine star alleles from sequence data[17–19]. For example, the tools Astrolabe and Stargazer, identifies patterns of SNVs and SVs in variant calling and alignment data and maps them to star alleles.

Over the past 11 years the AS has demonstrated clinical utility and gained acceptance as a tool to determine the phenotype from genotype, however, it has several limitations as appreciable variance exists within a given AS group[20]. A number of genetic factors have been proposed to explain the limited predictive ability of the AS. (1) Substrate specific effects make assigning a single score to each star allele difficult[21]. (2) There are variants that contribute to phenotype that have not yet been discovered, published, catalogued and assigned a star allele, so cannot be considered by the AS. (3) Star alleles are assigned a function by gene experts based on *in vivo* and *in vitro* evidence in order to be used to predict phenotype. Though computational methods could be used to predict function based on individual variants, to date no method to predict function for a haplotype has been embraced by the clinical community. However, there are many star alleles with conflicting, little, or no evidence of functional status in the published literature so function cannot be manually assigned. In fact, there are more than 57 star alleles in PharmVar with unknown function.

Any system that depends on star alleles may also have limited utility in diverse populations which have not been well studied. Since star allele definitions depend on what is submitted to and catalogued by PharmVar, bias in the ethnic populations represented by sequencing and genotyping studies will lead to bias in the frequencies of star alleles found in those populations. Indeed, this phenomenon has been documented in the clinical utility of genetic risk scores[22]. Many pharmacogenes have variants found at different frequencies depending on ancestry. Issues of ethnic bias in published studies, and therefore the catalogued star alleles, could be mitigated with a model that did not rely on curated literature and submitted haplotypes.

There exists a need to screen novel genotypes and haplotypes *in silico* in order to assess function. A computational approach to predict metabolic function would allow for screening of submitted haplotypes for which no *in vitro* testing has been done and provide insight to function which could be clinically useful for understudied populations where variation has not been well-studied or individuals with rare variations. An ideal computational approach would comprise the following items:

- It would be able to consider any possible coding variation in *CYP2D6*, such that rare or novel variants could be assessed as well as known variants.
- It would consider non-coding regions.
- It would provide substrate specific predictions.
- It would include information about structural variants.

No methods exist for predicting *CYP2D6* phenotype that perform all of these tasks. Methods for predicting variant deleteriousness are abundant, but these methods are often developed to be general purpose throughout the genome. Even when gene-specific methods exist, these methods are focused on prediction of the impact of single variants on function, rather than a set of variants in a genotype on organism-level function. *CYP2D6* is a complex gene with complex haplotypes and consequential genotypes and an optimal framework for predicting function should consider all these factors in unison.

Deep learning has emerged as a powerful tool that has revolutionized computer vision and has been used successfully in genetics[23]. It is most frequently applied in functional genetics for the prediction of motifs, such as transcription factor binding sites[24]. The power of deep learning in computer vision stems from the ability to allow the network to learn features that are important for prediction. This is primarily done using convolutional neural networks (CNNs). CNNs can learn spatial, structural, and sequential features, all of which are key features of the genome. CNNs have not been used for phenotype prediction, likely due to very limited data and the polygenic nature of most phenotypes. Despite the decline in cost of DNA sequencing and rise of large biobanks, genetic studies frequently suffer from small sample size. This is especially true in pharmacogenetics where identifying and phenotyping patients can be challenging. While deep learning may be an attractive solution for many problems, they require large amounts of data to successfully learn a new task, and thus our ability to apply neural networks to small data sets with little labeled data is limited.

We have developed a system, Hubble2D6, that predicts *CYP2D6* genotype and haplotype function. We present two models: a genotype-based activity predictor for predicting fine-grained drug metabolism of genotypes, and a haplotype classifier which can be used to screen haplotypes of unknown function. We demonstrate a method to build a computational model of a highly polymorphic gene with little labeled data by using simulations and unlabeled data, both labeled with the existing gold standard method. We show that creating a neural network model that emulates the output of the gold standard method has the flexibility to make predictions about previously unseen variants. We generate predictions for 57 *CYP2D6* star alleles of unknown function. We validate both models using *in vitro* studies from literature of 46 star alleles not previously seen by the models. We additionally perform *in silico* mutagenesis to assess the impact of every possible SNV within the *CYP2D6* locus.

Methods

Data

To train the genotype-based activity predictor we used two data sources, liver microsome data with measured metabolic activity and whole genome sequencing (WGS) without CYP2D6 metabolic measurements. We used 314 liver microsome samples from a prior study to train our model[14]. The liver microsome data was collected from two sites, 249 samples from St. Jude's Children's Research Hospital (SJCRH), and 65 samples from University of Washington (UW). Sequencing data was from the PGRNseq panel[25]. Metabolic activity was measured using two substrates: dextromethorphan and metoprolol. Additional data from 475 deeply sequenced genomes were used for weakly supervised learning from a separate WGS study[26]. Sequences from both studies had been aligned to hg19 and provided variant data in variant call format (VCF) files. Variants from WGS were extracted for the *CYP2D6* capture window on the PGRNseq panel. Star alleles, *CYP2D6* structural variants, and ASs were determined for all samples using Stargazer[17]. We removed 29 samples from the WGS data because they contained star alleles with unknown function.

Data Representation

We use a functional representation of genetic variants as input for our deep learning models (Fig. 1D). Variants are annotated with functional annotations and converted into a vector encoding of the nucleotide and its corresponding annotations. We fill in reference calls in the

VCFs to include every base in the capture window in order to represent the full gene sequence and annotate every nucleotide in the window (22:42521567-42528984, hg19).

We use annotations that are known to be important for protein function and gene expression. We use the following nine binarized annotations:

1. If the variant is in a coding region, as defined by RefSeq[27].
2. If it is rare in the population. Defined as allele frequency among all populations in gnoMAD < 0.05.
3. If it is deleterious. If it is a coding variant we use the ADME optimized framework defined by Zhou et al. If it is non-coding we use a majority vote of CADD, DANN, and FATHMM[28–31]. LOFTEE predictions of deleteriousness supercede the other methods, if available[32].
4. If it is an INDEL of any length. INDELs are reduced to the first nucleotide and given the INDEL annotation, so as to keep the length of each sequence the same.
5. If it is in a methylation mark, as defined by UCSC Genome Browser tracks wgEncodeHaibMethyl450Gm12878SitesRep1 and wgEncodeHaibMethylRrbsGm12878HaibSitesRep1[33,34].
6. If it is in a DNase hypersensitivity site. UCSC Genome Browser track wgEncodeAwgDnaseMasterSites.
7. If it is in a transcription factor binding site. UCSC Genome Browser tracks tfbsConsSites and wgEncodeRegTfbsClusteredV3.
8. If it is a known *CYP2D6* expression quantitative trait loci (eQTL) for any tissue in gTEX v6[35].

9. If it codes for a residue in the CYP2D6 active site where the substrate binds to the protein[36].

Sequences are annotated using Annovar for all annotations except LOFTEE, which is performed using VEP[37,38].

The annotated VCFs are divided into haplotypes for each sample and each haplotype sequence is one-hot encoded. Phasing of haplotypes is not performed. The resulting haplotype matrices are concatenated together with a zero matrix in between of length 50.

Model Building

Activity Predictor

Training

We developed a genotype-based activity predictor to predict the measured metabolic activity of liver microsomes for dextromethorphan and metoprolol substrates (Fig 1A). The model takes as input three items: (1) a genotype matrix of the full gene sequence with a functional representation, (2) a vector of structural variation (number of gene copies and counts of gene hybridization events), and (3) a binary variable indicating substrate (metoprolol or dextromethorphan). The model outputs a continuous variable as a prediction of metabolic activity. The model is a CNN that follows the basset architecture, with three convolutional layers and two fully connected layers[39]. Models were trained using Keras v2.2.4 with a Tensorflow v1.13.0 backend[40,41].

The genotype-based activity predictor was pretrained on simulated *CYP2D6* data (Fig 1C). Simulations were done by randomly selecting a pair of *CYP2D6* star alleles with known function (normal, decreased, or no function haplotypes) that do not have structural variants (neither hybridizations nor copy numbers are included) and constructing haplotypes with the variants associated with the star alleles. To introduce additional diverse training data, alternate alleles were sampled for variant sites not associated with any star allele following a uniform distribution with the probability of an alternate allele occurring equal to the population level alternate allele frequency published in gnoMAD[42]. We selected 20,000 genotypes of each AS (0, 0.5, 1, 1.5, 2), for a total of 100,000 simulated samples used in training, and an additional 20,000 total genotypes to use as a test set. No samples with an AS above 2 were selected because structural variants were not included in the pretraining of the model. The model was then trained to classify each genotype as its corresponding AS. After pretraining, the weights from the convolutional layers were transferred to a new network with the fully connected layers randomly initialized (Fig. 1F).

The model was then trained to predict the measured metabolic activity of the two substrates for each sample. Structural variants were determined by Stargazer and included in the model as a count of the total number of *CYP2D6* copies identified in the sample and a count of the hybridization events. These counts are appended to the vector that is output from the final convolutional layer so that they are included as input to the fully connected layers. We trained models through 5-fold cross validation, training on three folds, performing model selection using the fourth fold, and testing the model on the final fold. Data from SJCRH was used for training, and data from UW was used as a held out validation set. For each fold as many as 159 samples were included for training, 53 for validation, and 53 for test, and all 65 samples from

UW were held out for final validation. Data for both substrates is pooled together in the training set, so samples with measurements for both substrates are represented twice.

Weakly Supervised Learning

The training data were augmented with unlabeled WGS data in order to perform weakly supervised learning, specifically inaccurate supervision [43]. Labels were generated for the WGS data by training a linear regression model to predict the measured metabolic activity from the AS for SJCRH samples based on the three training folds, then predicting the metabolic activity of each WGS sample for the two substrates (Fig 1D). The data was then pooled with the labeled data for training of the genotype-based activity predictor. We train an ensemble of ten models for each fold, then take the mean of the ten models as the final prediction.

Model Evaluation

Models from each fold are evaluated using the coefficient of determination (R^2). We calculate the mean R^2 of the test folds and the mean R^2 of the models from each fold on the held out UW data.

As a baseline, we train a linear regression model to predict metabolic activity of both substrates from the AS. A linear model is fit using only the AS as a feature. A coefficient is learned for each AS bin, as in the following equation.

$$\hat{y} = \beta_1 AS_{0.5} + \beta_2 AS_1 + \beta_3 AS_{1.5} + \beta_4 AS_2 + \beta_5 AS_{2.5} + \beta_6 AS_3 + \epsilon$$

The highest AS observed in the training data is 3, hence the linear model learns coefficients up to an AS of 3. We train models using 5-fold cross-validation, with the same sample splits used in the genotype-based activity predictor.

In silico mutagenesis

We perform *in silico* mutagenesis to interpret the model weights for each variant and the model's ability to predict known deleterious variants. We create a new sequence corresponding to each nucleotide position in the *CYP2D6* capture window and each its possible alternate alleles, yielding 22,257 sequences. The generated sequences are homozygous for the alternate allele. The altered sequences are then passed through the genotype-based activity predictor and metabolic activity is predicted. We calculate the percent of baseline change for each variant sequence by dividing the predicted activity by the predicted activity for the reference sequence.

Star Allele Functional Classifier

Training

We train a star allele classifier by fine-tuning the genotype-based activity predictor with an added sigmoid activation layer to classify sequences as "normal function" or "reduced function". The model takes as input the genotype matrix and outputs the functional classification. We define reduced function as either "no function" or "decreased function" alleles. Increased function alleles are not included here, because presently increased function alleles are only defined by having multiple copies of a functional allele and copy number variation is not input into the star allele classifier.

We fine-tune the model on sequences constructed based on star allele definitions in PharmVar. We construct sequences for all known star alleles, then retrain the model using only star alleles and their suballeles that were observed in either the SJCRH liverbank or WGS data. The training set contains 15 star alleles, with 101 total suballeles.

Evaluation

We evaluate the model by predicting the function of the remaining 25 star alleles with known function that were excluded from the training process, and predict the function of 57 star alleles of unknown or uncertain function. We calculate the area under the receiver operator characteristic curve for the training and test groups.

We interpret the weights applied to each variant in the star alleles by the star allele classifier using DeepLIFT[44]. DeepLIFT compares the activiations of a neural network for a given sample against a reference sample, and outputs importance scores for each input feature. We run DeepLIFT on each star allele sequence with a *CYP2D6*1* reference sequence. This yields importance scores for each variant in each star allele that are different from the variants in *CYP2D6*1*.

Literature Validation

We validate the genotype-based activity predictor and star allele classifier using *in vitro* data from literature. We identified two sets of studies to use for validation: (1) a functional characterization of 49 *CYP2D6* star alleles performed using three substrates, and (2) eight *in vitro* studies of eleven *CYP2D6* star alleles discovered in Han Chinese subjects using ten substrates[45–54]. We exclude fourteen star alleles which were found in our training data, and evaluate on measurements for 46 star alleles that were not in our training data. We use the average metabolic activity (as a percent of *CYP2D6*1* activity) across all substrates included in the study for each of the star alleles as the measured metabolic activity.

We construct sequences containing the variants from each star allele, so that the genotype of each construct is homozygous for the star allele being tested (e.g. **24/*24*). We predict metabolic activity of dextromethorphan using the genotype-based activity predictor and compare the prediction against the mean metabolic activity for each *in vitro* tested star allele. Additionally, we predict normal vs reduced function using the star allele classifier. To format the prediction as a classification problem, we define star alleles with an activity greater than 50% of *CYP2D6*1* activity as "normal" and those with less than 50% as "reduced".

Methods Evaluation

We analyze the added components of our model by training a base model with one subtracted component at a time. We tested the effect of model pretraining, weakly supervised learning, and the functional variant representation. We calculate learning curves for models trained by dropping out one component at a time. We also calculate learning curves for models trained with all components and no components. To test the effect of pretraining the convolutional layers were randomly initialized rather than transferred pretrained model. To test the effect of weakly supervised learning, the model was trained using only liver microsome samples. To test the effect of the functional variant representation we removed the rows from the input matrix corresponding to the added annotations. We compare each learning curve to the learning curves of the full model and a model trained with no added components (no pretraining, no weakly supervised learning, and no functional variant representation).

We evaluate the contribution of the annotations included in the functional variant representation by training new models by training a new model with a single row removed for each annotation, resulting in thirteen models (four nucleotides, nine annotations). For each annotation, we calculate the percent decrease in R^2 in the resulting model.

Results

Activity Predictor

Predictions from the genotype-based activity predictor for each substrate and each collection site are shown in Fig. 2. A summary of the results compared to the AS is shown in Table 1. The genotype-based activity predictor performed similarly to the AS, with an R^2 of 0.71 for the dextromethorphan predictions on the UW samples, similar to 0.69 for the AS. Notably, the R^2 of the metoprolol predictions by the genotype-based activity predictor was lower than that of the AS for both collection sites. The pretrained model trained on simulated data predicted the AS of 20,000 held out simulations with 100% accuracy.

We perform *in silico* mutagenesis to interpret the model weights for different types of variants and the model's ability to predict known deleterious variants. The percent change in predicted activity from baseline for each variant can be seen in Figure 3. We show that the model has learned to predict variants known to be damaging to metabolic activity, even those that have not been previously seen by the model. Variants known to be deleterious in existing star alleles are annotated in Fig. 3D. Of the sixteen deleterious star allele associated variants the model had only previously seen four but predicts them to have a substantial negative impact on activity.

Star Allele Classifier

We trained a classification model to classify star alleles as normal function or reduced function by retraining the genotype-based activity predictor. In Figure 4 we show predictions for each star allele. We differentiate between star alleles that were in the training data, those that were not, and star alleles of unknown function. For both the training and test set we achieve an accuracy of 100% and an AUC of 1 in predicting normal vs reduced function.

We interpret the importance of the variants in each star allele to the resulting prediction of the star allele classifier using DeepLIFT using *CYP2D6*1* as a background (Fig. 5). Negative scores indicate variants that drove the prediction towards reduced function compared to the background, while positive scores indicate variants the model positively influence the prediction in relation to the background. We show that for haplotypes with known causal variants DeepLIFT attributes the largest weights to the causal variants (e.g. **4*, **10*, **19*).

We predicted function of all 57 star alleles with unknown function and find that we predict 31 of them to be reduced function and 26 to be normal function (Fig. 4). Of the star alleles predicted to be reduced function, four have variants that would likely be loss-of-function variants such as frameshift INDELs (e.g. **124*) or stop gain variants (e.g. **81*, **120*, **129*).

Literature Validation

We sought validation of our model through variants tested *in vitro* and published in literature. We identified one study that had published *in vitro* data for 49 star alleles with three substrates, and a set of eight papers that tested eleven star alleles identified in the Han Chinese population with ten substrates. We excluded star alleles that were present in our training data, leaving 46 star alleles for validation that had never been seen before by our models. We ran both the genotype-based activity predictor and the star allele classifier. We compare the dextromethorphan activity predictions from the genotype-based activity predictor and find that the predictions correlate with the measured data with an R^2 of 0.38. Using a cutoff of 50% of *CYP2D6*1* activity for normal vs reduced function, the star allele classifier achieves an AUC of 0.85 (Figure 5).

Method Evaluation

We analyzed the contribution of each of the techniques used in our training process, namely pretraining, weakly supervised learning, and the functional variant representation (Fig. 7A). Removing the function variant representation leads to the greatest increase in mean squared error, followed by weakly supervised learning, then pretraining. Combining all methods and evaluating the learning curve shows that the mean squared error (MSE) plateaus after 100 training samples. In the model with all labeled samples included, removing all components lead to a 36% increase in MSE, removing only the functional representation increased MSE by 22%, weakly supervised learning 13%, and pretraining 6%.

We evaluate the contribution of the selected annotations to the functional variant representation by creating new models that leave a single annotation out and calculating the percent decrease in R^2 (Fig 7B). The “deleterious” annotation leads to the greatest decrease in R^2 (5.5%), and the “rare” annotation affects the R^2 the least (0.6% decrease).

Discussion

Here we present Hubble2D6, a deep learning framework for the prediction of drug metabolic function of CYP2D6. We have constructed two models: 1) a genotype-based activity predictor that predicts substrate specific CYP2D6 metabolic activity for *CYP2D6* genotypes, and 2) a classification model that predicts whether a *CYP2D6* haplotype will be of normal function or reduced function. The two models serve two different purposes. The genotype-based activity predictor predicts a continuous enzymatic activity for combinations of haplotypes for either of the two substrates the model was trained on, dextromethorphan and metoprolol. The star allele classifier outputs normal vs reduced functional predictions of haplotypes which may be useful in the screening of haplotypes of unknown function. Both models take in the full *CYP2D6* gene sequence, including exons, introns, and the regions immediately upstream and downstream of the gene, as well as information about structural variants.

We devised a scheme for training a deep learning model to predict genotype and haplotype function with very little data. A frequent problem in genetic studies is limited availability of labeled data for making phenotype predictions and inferences about functional impact of variants, and deep learning models normally require vast amounts of data. We exploit the availability of a well performing baseline method for predicting *CYP2D6* function, the AS. The AS is a thoughtfully devised tool that performs well in predicting the function of *CYP2D6* genotype. We utilize the AS at multiple points in our method. We pretrain the network on simulated data labeled with the AS which yields a network with the AS rules encoded, then, while fine tuning the model we include unlabeled *CYP2D6* sequences from WGS data labeled using the AS. We show that by pretraining a network on simulated data and including labeled using the AS, we can train a CNN that performs as well as the human curated baseline with the ability to assess novel sequences for which functional assignments do not yet exist. Pretraining provides the neural network with more than 300x the amount of data available in the training set, allowing it to learn properties of the genes important for predicting activity. Related approaches have been shown to be effective in computer vision[55,56]. Similar pretraining and weakly supervised learning approaches may be used for other genes where a baseline method exists for predicting function.

Our approach allows for the input of arbitrary sequences and therefore can make predictions about the function of all possible variants and combinations of variants in *CYP2D6*. Although our genotype-based activity predictor does not outperform the AS in terms of variance explained, it provides a huge advantage because it allows for predictions of novel variants and haplotypes. The output from the genotype-based activity predictor for the star alleles derived from literature studies had an R^2 of only 0.38, but since many of these had unknown function the AS would not have assigned a functional prediction. New variants and haplotypes are frequently being discovered, so the need to assess all variants and combinations of variants is an important component of the model. We demonstrate this capability through *in silico* mutagenesis, mutating every base in the gene sequence to its alternate allele (Fig. 3), and by predicting the function of all star alleles of unknown function (Fig. 4).

We predicted the function of all 57 star alleles with unknown function using the star allele classifier and found that 31 of them are predicted to be reduced function alleles. Although the function remains unknown, four of the unknown star alleles harbor loss-of-function variants that would typically lead to a non-functional protein. Functional predictions for these star alleles with unknown function may increase the clinical utility of the AS. *In silico* prediction of haplotype function may be able to provide additional guidance to those wishing to use the AS who encounter star alleles of unknown function.

We evaluate our models on *in vitro* data for haplotypes that have were previously unseen by the model from two existing studies. Many of these are haplotypes for which no official functional designation is available on PharmVar. Additional variability may come from study design parameters, such as the expression system, rather than a genetic contribution to changes in haplotype function. Finally, the haplotypes from Qian *et al* are discovered in Han Chinese population and our model was developed on samples that are primarily from donors with European ancestry. Despite these challenges, we find that the star allele classifier performs moderately well in assessing the function of the measured haplotypes. In order to assess the performance of the classification model we set a cutoff of 50% of wild-type activity for determining “normal” vs “reduced” function. Although there is no official cutoff between normal and decreased function alleles we used this cutoff in order to assess model performance. We find that the model performs well differentiating between normal and reduced function alleles, with an AUC of 0.84. An ideal model would predict normal, decreased, and no function labels for haplotypes, however, we were unable to train a model that had high accuracy differentiating between the three functional groups. This is likely due to the relatively small number of examples of decreased function alleles in the training set. Additionally, there may be alleles with increased function, but there are presently no star alleles with increased function that is not caused by increased copy number. Differentiating between normal function and reduced function may be sufficient for triaging star alleles and giving an indication of function.

We examined the features that most contribute to prediction of phenotype. In the star alleles discovered by Qian et al there are three star alleles containing nucleotide polymorphisms

(100C>T) that lead to a P34S amino acid change (**87*, **94*, **95*). This variant is frequently found among *CYP2D6* star alleles and is the core SNV of **10*, which is a decreased function star allele. All three P34S containing star alleles are predicted to be of reduced function by the star allele classifier, however **94* has the highest average activity among all Qian et al star alleles with an average activity of 89% of wild type (Fig. 6A). DeepLIFT reveals that the model assigned a negative weight to the P34S variant in **94*, which likely lead to its classification as reduced function (Fig. 5). This demonstrates that the model has learned to predict all P34S containing star alleles as reduced function.

We show that the star allele classifier can be interpreted, which may enhance clinical utility (Fig. 6). Using DeepLIFT, we calculate importance scores for each variant in the star allele which provides insight to which variants the model considered to have the biggest impact on function. For example, we see that DeepLIFT assigns the largest negative scores to the splicing defect in **4* and the frameshift INDEL in **19*. This functionality may be critical when evaluating the accuracy of the predictions.

There are several limitations to our *CYP2D6* predictive models. First, although we have worked to overcome the small amount of labeled data, the minimal data we used could lead to a model that has not learned all possible effects of variants on enzymatic function. The data is limited by natural variation that is observed in the human population, and even further limited to the samples used in this study which are predominantly European. Mutagenesis methods could yield richer data to learn a more robust model of metabolic activity[57]. Second, since we only consider a narrow window around *CYP2D6* we miss opportunities for more distal effects on gene expression, such as the long-range enhancer associated with the *2 haplotype[58–60]. These distal regulatory effects are not captured by the current model, and thus our ability to fully explain the variance in enzymatic activity is limited to the capture region on the PGRNseq platform. Third, although we find that the inclusion of structural variants does improve prediction, we believe this could be further improved. The copy number and occurrence of hybridization events that we include in the model is not associated with either of the input haplotypes. If we were able to assign copy number or fusion events with phased data this could improve the predictions in the genotype-based activity predictor. Fourth, samples were not phased prior to being input to the model. Phased input was tested but performed worse than inputting unphased data. Phase should be important for predicting function, so this finding is counterintuitive. Finally, there are factors that affect *CYP2D6* metabolic activity outside the gene sequence[61–63].

In conclusion, we have created two models for the prediction of *CYP2D6* metabolic activity from sequence data that could expand our ability to predict metabolic activity from sequence. Our approach has a variety of unique features, which include pretraining a convolutional neural network on simulated sequence data, weakly supervised learning on unlabeled data, and a functional representation of genetic data that includes annotations. These methods could be applied to other genes for which single gene prediction models would be useful. We envision that our predictions about *CYP2D6* haplotype function may be used to triage star alleles for which *in vitro* testing has not yet been performed.

Acknowledgements

G.M. is supported by the Big Data to Knowledge (BD2K) from the National Institutes of Health (T32 LM012409). E.L.W. and R.D. are supported by the Northwest Alaska-Pharmacogenomics Research Network (NWA-PGRN) (P01GM116691). K.S., M.W.C., and R.B.A are supported by

NIH/National Institute of General Medical Sciences PharmGKB resource, (R24GM61374). We thank Kenneth Thummel at the University of Washington and Erin Schuetz at St. Jude Children's Research Hospital for providing the liver bank resources to conduct this work. We also thank Phil Tsao for providing the WGS data used in this study.

References

1. Relling, M. V. & Klein, T. E. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clin. Pharmacol. Ther.* **89**, 464–467 (2011).
2. Whirl-Carrillo, M. *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **92**, 414–417 (2012).
3. Van Driest, S. L. *et al.* Clinically actionable genotypes among 10,000 patients with preemptive pharmacogenomic testing. *Clin. Pharmacol. Ther.* **95**, 423–431 (2014).
4. Chanfreau-Coffinier, C. *et al.* Projected Prevalence of Actionable Pharmacogenetic Variants and Level A Drugs Prescribed Among US Veterans Health Administration Pharmacy Users. *JAMA Netw Open* **2**, e195345 (2019).
5. Reisberg, S. *et al.* Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions. *Genet. Med.* (2018). doi:10.1038/s41436-018-0337-5
6. Zhou, S.-F. Polymorphism of human cytochrome P450 2D6 and its clinical significance: Part I. *Clin. Pharmacokinet.* **48**, 689–723 (2009).
7. Zhou, S.-F. Polymorphism of human cytochrome P450 2D6 and its clinical significance: part II. *Clin. Pharmacokinet.* **48**, 761–804 (2009).
8. Gaedigk, A. Complexities of CYP2D6 gene analysis and interpretation. *Int. Rev. Psychiatry*

25, 534–553 (2013).

9. 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).
10. Gaedigk, A. *et al.* The Evolution of PharmVar. *Clin. Pharmacol. Ther.* **105**, 29–32 (2019).
11. Gaedigk, A., Sangkuhl, K., Whirl-Carrillo, M., Klein, T. & Leeder, J. S. Prediction of CYP2D6 phenotype from genotype across world populations. *Genet. Med.* **19**, 69–76 (2017).
12. Del Tredici, A. L. *et al.* Frequency of CYP2D6 Alleles Including Structural Variants in the United States. *Front. Pharmacol.* **9**, 305 (2018).
13. Black, J. L., 3rd, Walker, D. L., O'Kane, D. J. & Harmandayan, M. Frequency of undetected CYP2D6 hybrid genes in clinical samples: impact on phenotype prediction. *Drug Metab. Dispos.* **40**, 111–119 (2012).
14. Rachel Dalton, Seung-been Lee, Katrina Claw, Bhagwat Prasad, Brian R. Phillips, Danny D. Shen, Lai Hong Wong, Mitch Fade, Matthew G. McDonald, Maitreya J. Dunham, Douglas M. Fowler, Allan E. Rettie, Erin Schuetz, Andrea Gaedigk, Timothy A. Thornton, Deborah A. Nickerson, Kenneth E. Thummel, Erica L. Woodahl. Interrogation of CYP2D6 structural variant alleles increases the association between CYP2D6 genotype and CYP2D6-mediated metabolic activity. *Manuscript in preparation.* (2019).
15. J Marcalus, S. & Bristow-Marcalus, S. Combating opioid addiction and abuse--2 ways to effectively intervene in the cycle of addiction through pharmacogenomics. *J. Am. Pharm. Assoc.* (2019). doi:10.1016/j.japh.2019.04.016
16. Gaedigk, A. *et al.* The CYP2D6 activity score: translating genotype information into a qualitative measure of phenotype. *Clin. Pharmacol. Ther.* **83**, 234–242 (2008).
17. Lee, S.-B. *et al.* Stargazer: a software tool for calling star alleles from next-generation

sequencing data using CYP2D6 as a model. *Genet. Med.* **21**, 361–372 (2019).

18. Numanagić, I. *et al.* Cypiripi: exact genotyping of CYP2D6 using high-throughput sequencing data. *Bioinformatics* **31**, i27–34 (2015).
19. Twist, G. P. *et al.* Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, CYP2D6, from whole-genome sequences. *NPJ Genom Med* **1**, 15007 (2016).
20. Gaedigk, A., Dinh, J. C., Jeong, H., Prasad, B. & Leeder, J. S. Ten Years' Experience with the CYP2D6 Activity Score: A Perspective on Future Investigations to Improve Clinical Predictions for Precision Therapeutics. *J Pers Med* **8**, (2018).
21. Bogni, A. *et al.* Substrate specific metabolism by polymorphic cytochrome P450 2D6 alleles. *Toxicol. In Vitro* **19**, 621–629 (2005).
22. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
23. Zou, J. *et al.* A primer on deep learning in genomics. *Nat. Genet.* **51**, 12–18 (2019).
24. Park, Y. & Kellis, M. Deep learning for regulatory genomics. *Nat. Biotechnol.* **33**, 825–826 (2015).
25. Gordon, A. S. *et al.* PGRNseq: a targeted capture sequencing panel for pharmacogenetic research and implementation. *Pharmacogenet. Genomics* **26**, 161–168 (2016).
26. Pan, C. *et al.* Cloud-based interactive analytics for terabytes of genomic variants data. *Bioinformatics* **33**, 3709–3715 (2017).
27. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).
28. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**,

D886–D894 (2019).

29. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).
30. Shihab, H. A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536–1543 (2015).
31. Zhou, Y., Mkrtchian, S., Kumondai, M., Hiratsuka, M. & Lauschke, V. M. An optimized prediction framework to assess the functional impact of pharmacogenetic variants. *Pharmacogenomics J.* **19**, 115–126 (2019).
32. Karczewski, K. J. LOFTEE (Loss-Of-Function Transcript Effect Estimator). (2015).
33. Rosenbloom, K. R. *et al.* ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.* **41**, D56–63 (2013).
34. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
35. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
36. Zhou, S. *Cytochrome P450 2D6: Structure, Function, Regulation and Polymorphism*. (CRC Press, 2016).
37. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
38. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
39. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
40. Chollet, F. & Others. Keras. (2015). Available at: https://keras.io.
41. Martín Abadi *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous

Systems. (2015).

42. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019). doi:10.1101/531210

43. Zhou, Z.-H. A brief introduction to weakly supervised learning. *Natl Sci Rev* **5**, 44–53 (2018).

44. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arXiv [cs.CV]* (2017).

45. Muroi, Y. *et al.* Functional characterization of wild-type and 49 CYP2D6 allelic variants for N-desmethyltamoxifen 4-hydroxylation activity. *Drug Metab. Pharmacokinet.* **29**, 360–366 (2014).

46. Dai, D.-P. *et al.* In vitro functional assessment of 22 newly identified CYP2D6 allelic variants in the Chinese population. *Basic Clin. Pharmacol. Toxicol.* **117**, 39–43 (2015).

47. Cai, J. *et al.* Effects of 22 Novel CYP2D6 Variants Found in the Chinese Population on the Bufuralol and Dextromethorphan Metabolisms In Vitro. *Basic Clin. Pharmacol. Toxicol.* **118**, 190–199 (2016).

48. Liang, B. *et al.* Effect of 24 Cytochrome P450 2D6 Variants Found in the Chinese Population on Atomoxetine Metabolism in vitro. *Pharmacology* **97**, 78–83 (2016).

49. Wang, Z.-H. *et al.* Effects of 24 CYP2D6 Variants Found in the Chinese Population on the Metabolism of Risperidone. *Pharmacology* **96**, 290–295 (2015).

50. Zhan, Y.-Y. *et al.* Effect of CYP2D6 variants on venlafaxine metabolism in vitro. *Xenobiotica* **46**, 424–429 (2016).

51. Hu, X.-X. *et al.* Effect of CYP2D6 genetic polymorphism on the metabolism of citalopram in vitro. *Drug Metab. Pharmacokinet.* **31**, 133–138 (2016).

52. Weng, Q. *et al.* Effect of 24 cytochrome P450 2D6 variants found in the Chinese population on the N-demethylation of amitriptyline in vitro. *Pharm. Biol.* **54**, 2475–2479 (2016).

53. Hu, X.-X. *et al.* Functional characterization of 22 novel CYP2D6 variants for the metabolism of Tamoxifen. *J. Pharm. Pharmacol.* **68**, 819–825 (2016).

54. Qian, J.-C. *et al.* Genetic variations of human CYP2D6 in the Chinese Han population. *Pharmacogenomics* **14**, 1731–1743 (2013).

55. Mahajan, D. *et al.* Exploring the Limits of Weakly Supervised Pretraining. *arXiv [cs.CV]* (2018).

56. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? in *Advances in Neural Information Processing Systems 27* (eds. Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q.) 3320–3328 (Curran Associates, Inc., 2014).

57. Hess, G. T. *et al.* Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat. Methods* **13**, 1036–1042 (2016).

58. Wang, D. *et al.* Common CYP2D6 polymorphisms affecting alternative splicing and transcription: long-range haplotypes with two regulatory variants modulate CYP2D6 activity. *Hum. Mol. Genet.* **23**, 268–278 (2014).

59. Ray, B., Ozcagli, E., Sadee, W. & Wang, D. CYP2D6 haplotypes with enhancer single-nucleotide polymorphism rs5758550 and rs16947 (*2 allele): implications for CYP2D6 genotyping panels. *Pharmacogenet. Genomics* **29**, 39–47 (2019).

60. Wang, D., Papp, A. C. & Sun, X. Functional characterization of CYP2D6 enhancer polymorphisms. *Hum. Mol. Genet.* **24**, 1556–1562 (2015).

61. Sandee, D. *et al.* Effects of genetic variants of human P450 oxidoreductase on catalysis by CYP2D6 in vitro. *Pharmacogenet. Genomics* **20**, 677–686 (2010).

62. Crewe, H. K., Lennard, M. S., Tucker, G. T., Woods, F. R. & Haddock, R. E. The effect of selective serotonin re-uptake inhibitors on cytochrome P4502D6 (CYP2D6) activity in human liver microsomes. *Br. J. Clin. Pharmacol.* **58**, S744–S747 (2004).

63. Wadelius, M., Darj, E., Frenne, G. & Rane, A. Induction of CYP2D6 in pregnancy. *Clin. Pharmacol. Ther.* **62**, 400–407 (1997).

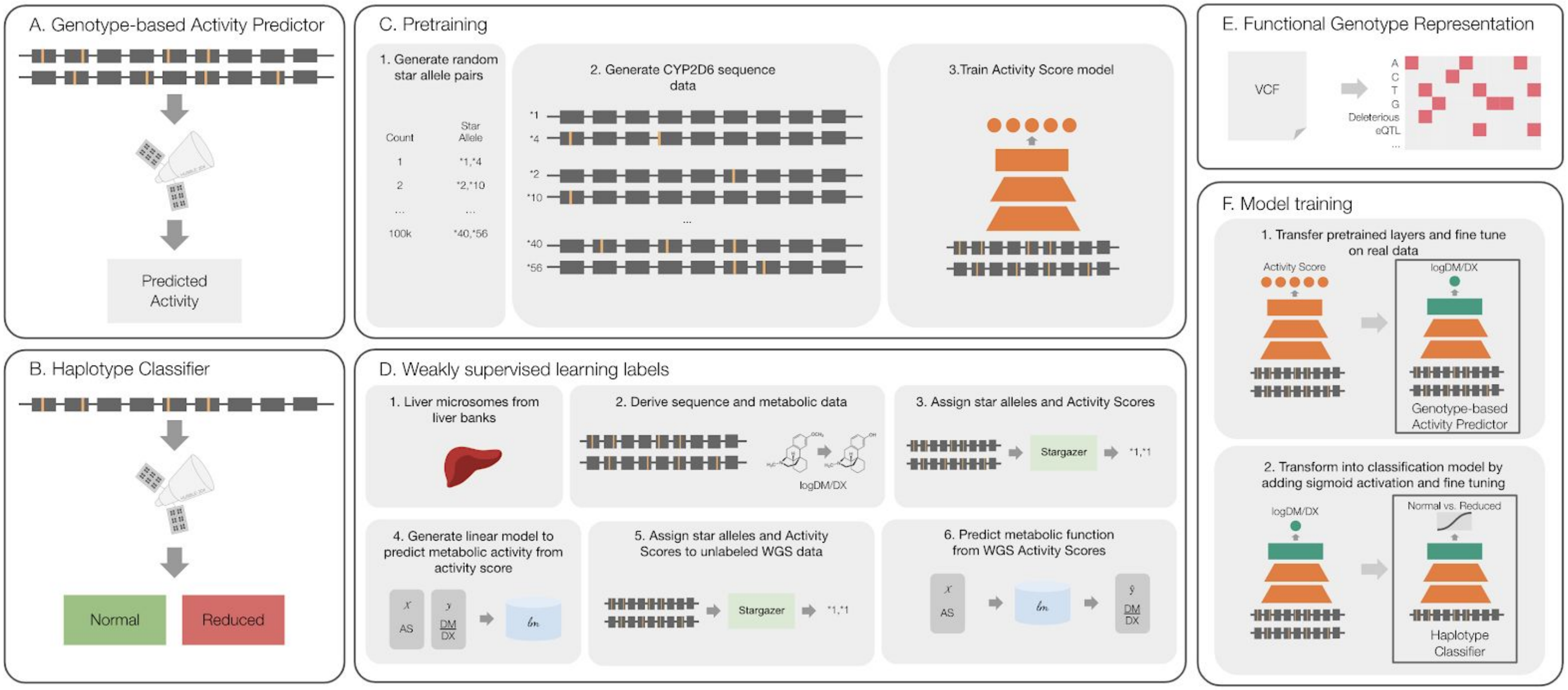
A. Genotype-based Activity Predictor
Predicted Activity
B. Haplotype Classifier
Normal
Reduced
C. Pretraining
1. Generate random star allele pairs
Count
Star Allele
1
*1,*4
2
*2,*10
...
100k
*40,*56
2. Generate CYP2D6 sequence data
*1
*4
*2
*10
*40
*56
3.Train Activity Score model
D. Weakly supervised learning labels
1. Liver microsomes from liver banks
2. Derive sequence and metabolic data
logDM/DX
3. Assign star alleles and Activity Scores
Stargazer
*1,*1
4. Generate linear model to predict metabolic activity from activity score
x
AS
y
DM
DX
lm
5. Assign star alleles and Activity Scores to unlabeled WGS data
Stargazer
*1,*1
6. Predict metabolic function from WGS Activity Scores
x
AS
lm
ŷ
DM
DX
E. Functional Genotype Representation
VCF
A
C
T
G
Deleterious
eQTL
...
F. Model training
1. Transfer pretrained layers and fine tune on real data
Activity Score
logDM/DX
Genotype-based Activity Predictor
2. Transform into classification model by adding sigmoid activation and fine tuning
logDM/DX
Normal vs. Reduced
Haplotype Classifier

Figure 1. Prediction workflow and model components. Hubble comprises two predictive models, a genotype-based activity predictor (A), and a haplotype classifier (B). We utilize three methods to address the challenge of data scarcity. First, we pretrain the model using simulated data (C), second we use weakly supervised learning to enhance our training data (D), and third, we use a functional variant representation, where each variant is represented as a vector of its nucleotide and annotations (E). To train the genotype-based activity predictor, we transfer the convolutional layers from the pretrained model to a new model and train it to predict the measured metabolic activity of liver microsomes (F, top). Then, to develop the haplotype classifier we add a sigmoid activation to the genotype-based activity predictor and fine-tune the model on simulated star allele sequences to predict normal function or reduced function (either decreased or no function alleles), (F, bottom).

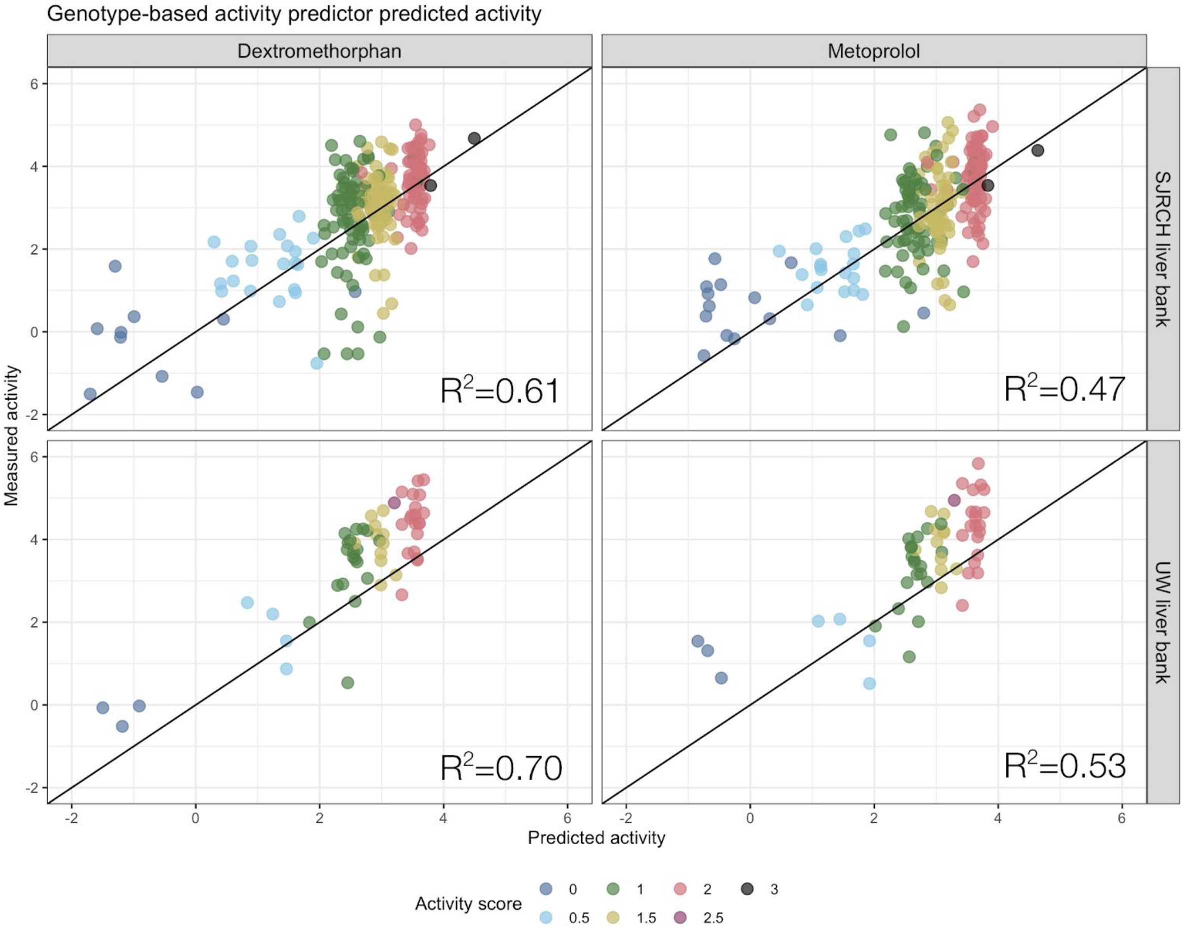

Figure 2. Metabolic activity predictions for the genotype-based activity predictor for each of the two substrates for samples from both liver banks. The genotype-based activity predictor was trained on data from SJCRH liverbank through 5-fold cross-validation and evaluated on samples from the UW liverbank. The top row shows predictions for the held out test for each of the five folds, the bottom row shows predictions for the samples from the UW liverbank. The left column shows predictions for dextromethorphan metabolism, the right column shows predictions for metoprolol metabolism. The colors indicate the AS for each sample.

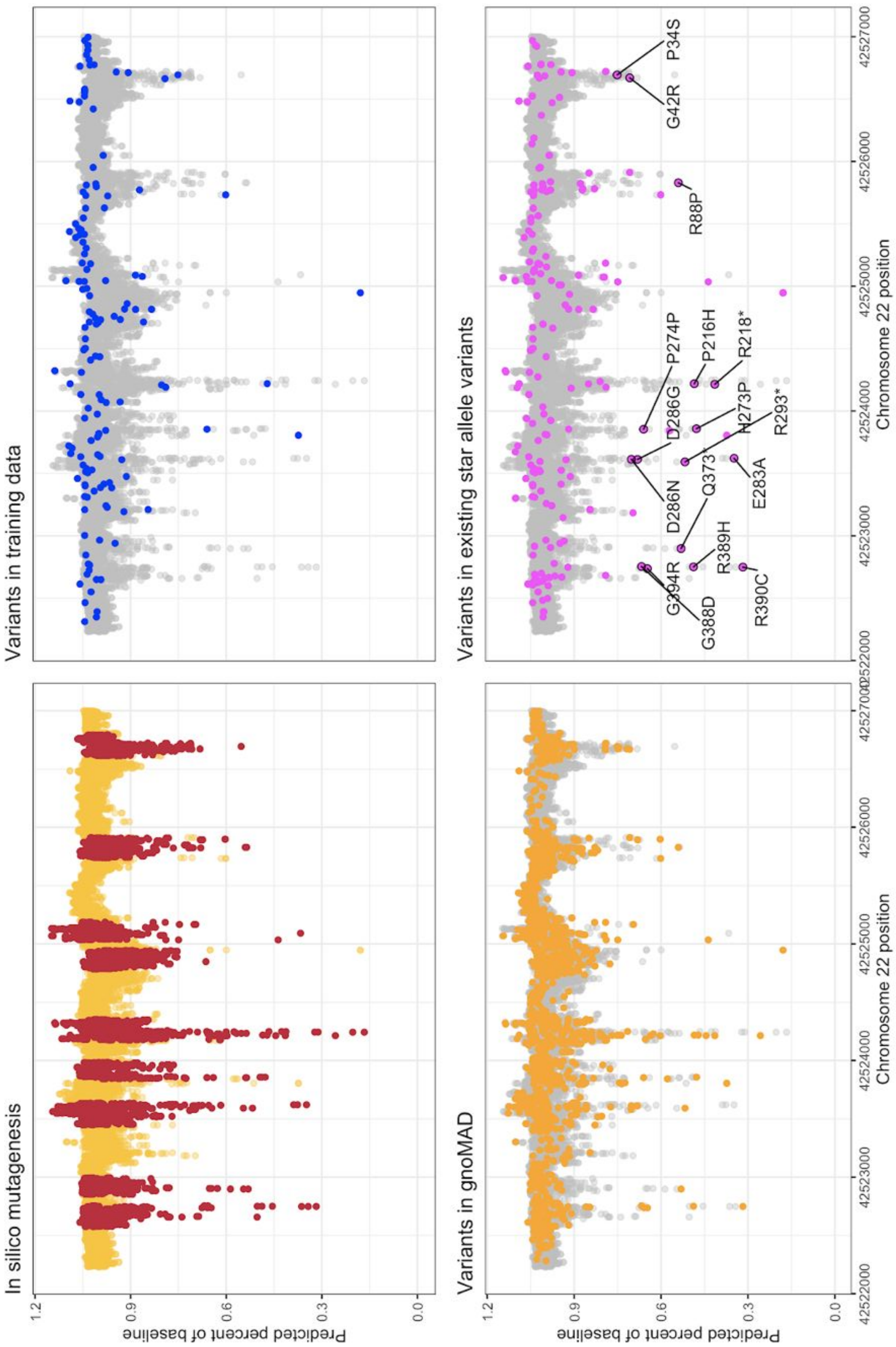
In silico mutagenesis
Variants in training data
Variants in gnoMAD
Variants in existing star allele variants
Predicted percent of baseline
1.2
0.9
0.6
0.3
0.0
Chromosome 22 position
42522000
42523000
42524000
42525000
42526000
42527000
G394R
G388D
R389H
R390C
D286N
Q373*
E283A
D286G
H273P
R293*
P274P
P216H
R218*
R88P
G42R
P34S

Figure 3. Figure 3. *In silico* mutagenesis of CYP2D6. In order to interpret the model we mutate every base in CYP2D6 to each possible alternate allele, yielding 21,000 sequences each with a SNV from the reference sequence. Each subplot here shows the same data, with the x-axis indicating the location of the SNV, and the y-axis indicating the predicted change in activity from wild type. The top left plot shows exons colored in red and noncoding variants colored in yellow. The top right plot highlights all variants that were observed in the training data in blue, variants not previously seen by the model are in gray. The bottom left plot displays variants in orange that have been observed in gnoMAD, previously unobserved variants are shown in gray. The bottom right plot shows variants that are in existing star alleles in magenta, with several of the known loss-of-function amino acid changes annotated. Variants not in existing star alleles are shown in gray.

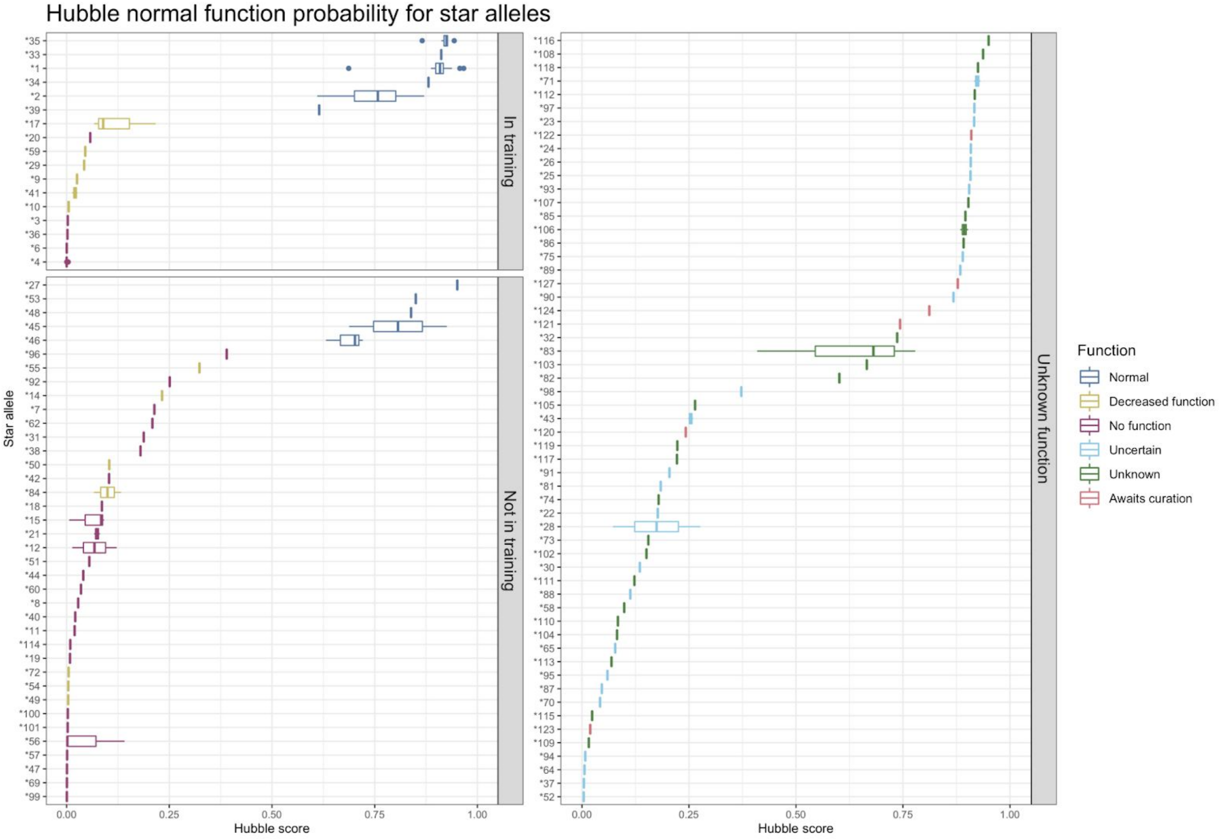

Figure 4 Star allele classifier predictions. Here we show the scores output from the star allele classifier for each star allele sequence. Star alleles are shown along the y-axis, with boxes differentiating whether the star allele was in the training set, the test set, or if it is of unknown function. Along the x-axis is the score for each star allele. Box plots are shown for star alleles with more than one sub allele. Scores greater than 0.5 correspond to predictions of normal function, and a score less than 0.5 indicates a prediction of reduced function. Star alleles are sorted by the median score from the model.

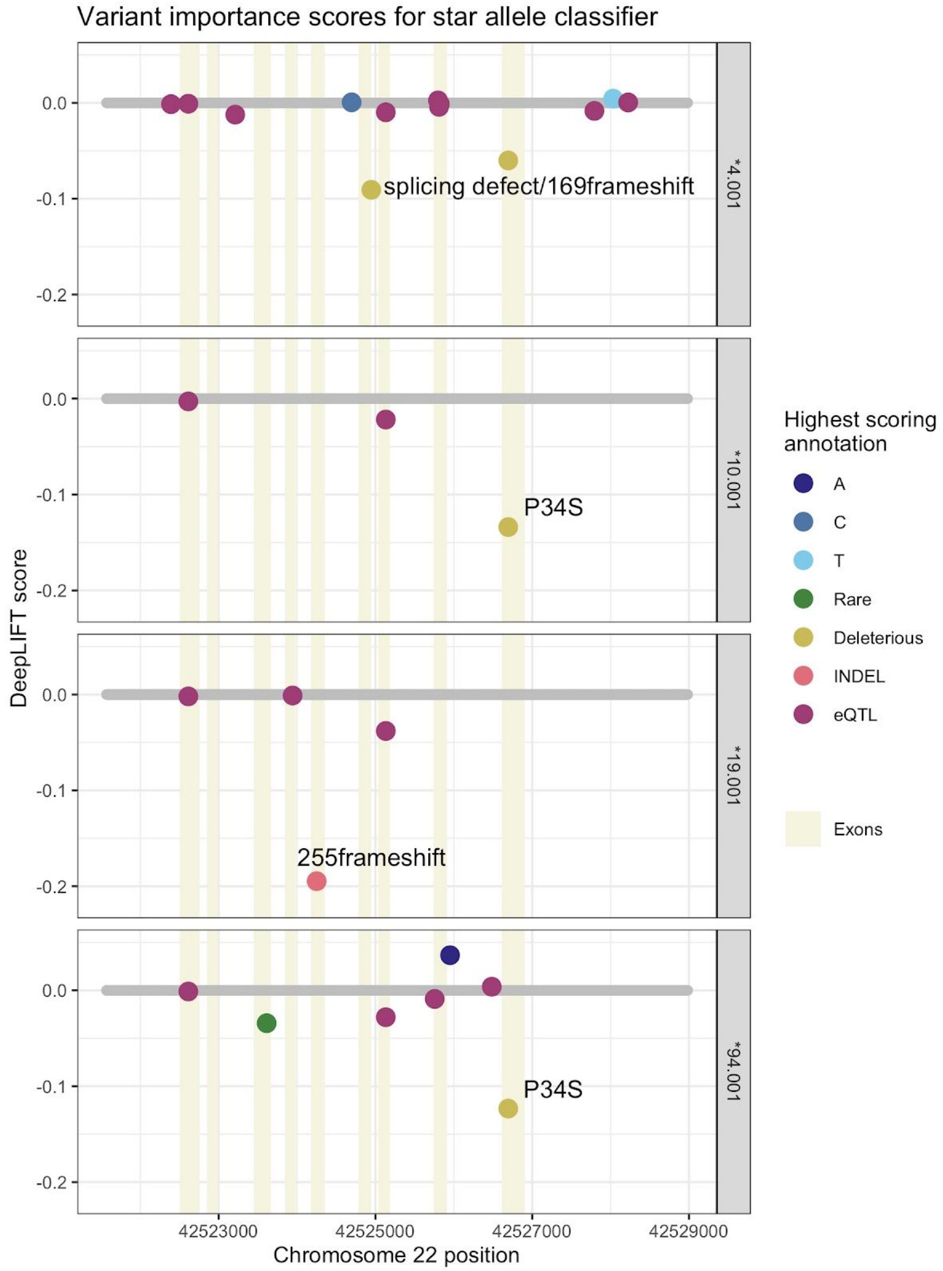

Figure 5. Variant importance scores for selected star alleles. Here we show variant importance scores from DeepLIFT that serve as a means to interpret the star allele classifier. The x-axis indicates the position of the variant in *CYP2D6*, the y-axis indicates the relative importance score. Negative scores indicate that the variant led the model to predict reduced function compared to the background allele, *CYP2D6*1*. The colors of each point represent the nucleotide or annotation that received the largest weight from the model, in terms of absolute value. Exons are shown in beige as a reference. The variant with the largest absolute weights are annotated with either the amino acid change they cause or the variant type.

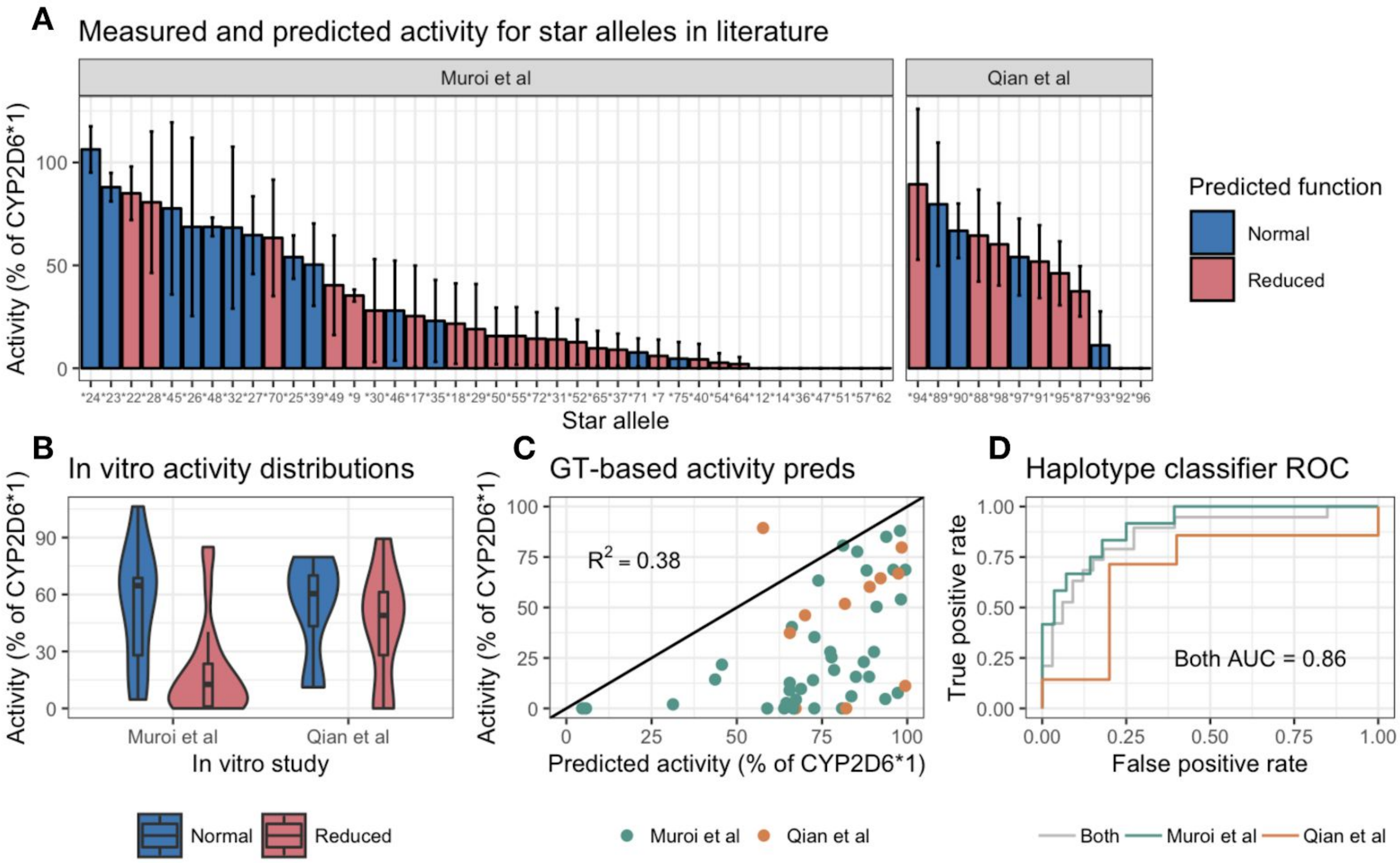
A
Measured and predicted activity for star alleles in literature
Muroi et al
Qian et al
Activity (% of CYP2D6*1)
100
50
0
*24*23*22*28*45*26*48*32*27*70*25*39*49 *9 *30*46*17*35*18*29*50*55*72*31*52*65*37*71 *7 *75*40*54*64*12*14*36*47*51*57*62
*94*89*90*88*98*97*91*95*87*93*92*96
Star allele
Predicted function
Normal
Reduced
B
In vitro activity distributions
Activity (% of CYP2D6*1)
90
60
30
0
Muroi et al
Qian et al
In vitro study
Normal
Reduced
C
GT-based activity preds
Activity (% of CYP2D6*1)
100
75
50
25
0
$R^2 = 0.38$
0
25
50
75
100
Predicted activity (% of CYP2D6*1)
Muroi et al
Qian et al
D
Haplotype classifier ROC
True positive rate
1.00
0.75
0.50
0.25
0.00
Both AUC = 0.86
0.00
0.25
0.50
0.75
1.00
False positive rate
Both
Muroi et al
Qian et al

Figure 6. Measured vs predicted activity for star alleles of unknown function with *in vitro* data. Here we show predictions of 26 star alleles that do not have functional assignment but do have *in vitro* metabolic data available. In the top plot we display a bar chart of the measured activity for each of the star alleles, with star alleles predicted by the star allele classifier to be normal in blue, and predicted to be of reduced function in red. The bottom left plot displays the same data as the top plot as violin plots to highlight the differences in the distributions between studies and between predictions from the star allele classifier. The bottom middle plot shows predictions from the genotype-based activity predictor for dextromethorphan metabolism for each of the 26 star alleles. The points are colored by which study they came from, with star alleles from Muroi et al in green and star alleles from Qian et al in orange. The R^2 was calculated on the star alleles from both sources combined. The bottom right plot shows a receiver operator characteristic curve for the star allele classifier for each of the star allele groups, with the star alleles from Muroi et al in green, Qian et al in orange, and both sets of samples grouped together in gray. To make this into a classification problem for we define samples with greater than 50% of CYP2D6*1 activity as "normal" and those with less than 50% as "reduced". The AUC was calculated on the star alleles from both sources combined.

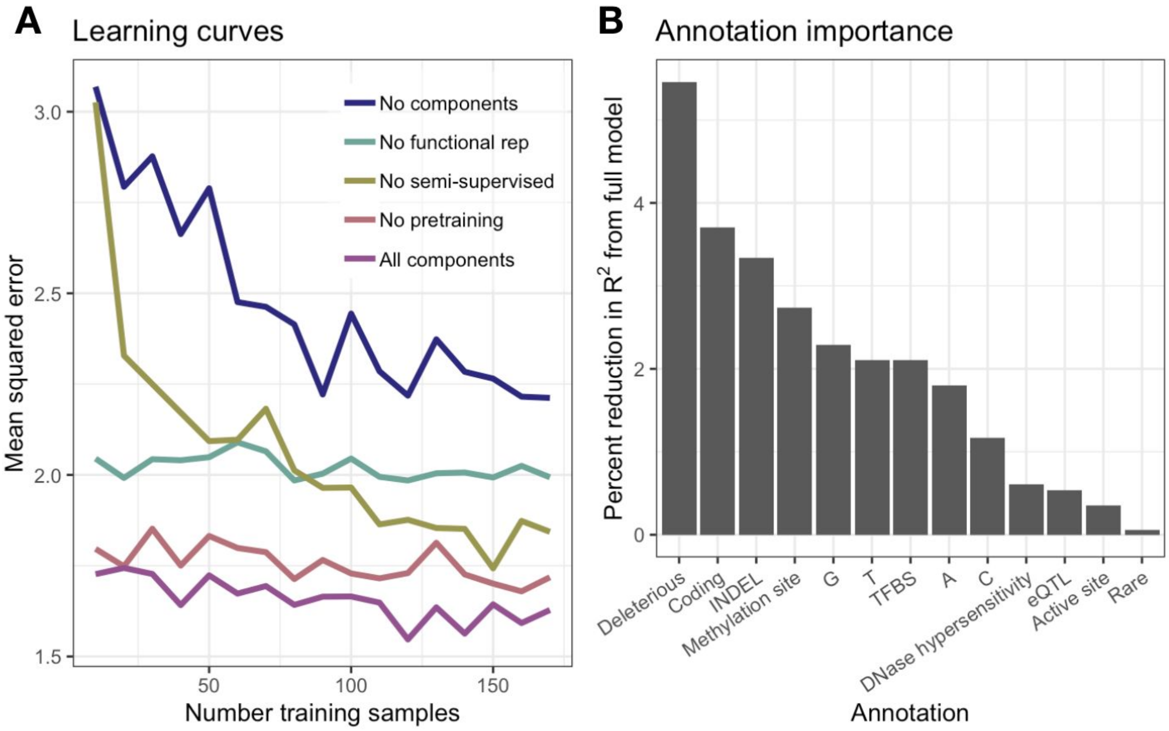

Figure 7. Method evaluation. Here we show an evaluation of the methods used to compensate for data scarcity. In part A, we show learning curves for models trained dropping out one component at a time. For example, the “No pretraining” line represents the learning curve for a model trained with weakly supervised learning and a functional variant representation, but no pretraining. We also show learning curves for models trained with all components and no components. In part B, we show an interpretation of annotation importance in the functional representation. Each bar represents an annotation included in the representation and the y-axis indicates the percent decrease in R^2 from the full model for a model trained without that particular annotation. Here, the height is positively correlated with annotation importance to the model.

Table 1. Genotype-based activity predictor results. In this table we show the results for both substrates of the Genotype-based activity predictor trained through 5-fold cross-validation on the liver microsomes from the SJCRH liverbank, and evaluated on the UW liverbank samples which were held out during training. We compare against the variance explained by a linear regression on the AS with the same 5-fold cross validation scheme. The R^2 values presented for the SJCRH data were the average value on the held-out test group from each fold. The R^2 values for the UW samples are the mean from the five models trained through cross-validation. The bolded values indicate the maximum R^2 achieved for each drug in each dataset.

		SJCRH Test Data R^2		UW Validation Data R^2	
Substrate	Method	Mean	SD	Mean	SD
Dextromethorphan	AS	**0.63**	0.09	0.69	0.01
	Hubble2D6	0.61	0.07	**0.71**	0.01
Metoprolol	AS	**0.49**	0.12	**0.57**	0.01
	Hubble2D6	0.48	0.13	0.54	0.01