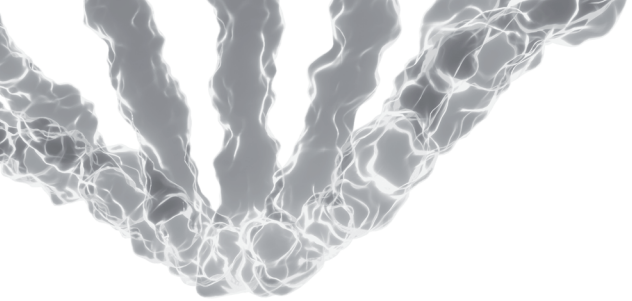
Towards solving the missing heritability in pharmacogenomics

Lee, M. van der



CHAPTER 5

Towards predicting CYP2D6-mediated variable drug response from *CYP2D6* gene sequencing data

Maaike van der Lee, William G. Allard, Rolf H.A.M Vossen, Renée F. Baak-Pablo, Roberta Menafra, Birgit A.L.M. Deiman, Maarten J. Deenen, Patrick Neven, Inger Johansson, Stefano Gastaldello, Magnus Ingelman-Sundberg, Henk-Jan Guchelaar, Jesse J. Swen, Seyed Yahya Anvar

Science Translational Medicine. 2021 Jul 21;13(603):eabf3637

Abstract

Pharmacogenomics is a key component of personalized medicine. It promises a safer and more effective drug treatment by individualizing the choice of drug and dose based on an individual's genetic profile. In clinical practice, genetic biomarkers are being used to categorize patients into predefined *-alleles to predict CYP450 enzyme activity and adjust drug dosages accordingly. Yet, this approach has substantial limitations as it leaves a large part of variability in drug response unexplained. Here, we present a proof-of-concept approach and introduce a continuous scale (instead of categorical) assignments to predict metabolic enzyme activity. We used the full *CYP2D6* gene sequence as obtained with long-read amplicon-based sequencing and Cytochrome P450 (CYP) 2D6-mediated tamoxifen metabolism data from a prospective study of 561 patients with breast cancer to train a neural network. The model explained 79% of the interindividual variability in CYP2D6 activity compared to 54% with the conventional *-allele approach and assigned accurate enzyme activities to known alleles (activity matched previously reported effects) and predicted the activity of previously uncharacterized combinations of variants. The results were replicated in an independent cohort of tamoxifen-treated patients (model R^2-adjusted = 0.66 *vs* *-allele R^2-adjusted = 0.35) and a cohort of patients treated with the CYP2D6 substrate venlafaxine (model R^2-adjusted = 0.64 *vs* *-allele R^2-adjusted = 0.55). Moreover, human embryonic kidney cells were used to confirm the effect of five variants in in vitro functional assays measuring the metabolism of the CYP2D6 substrate bufuralol. These results demonstrated the advantage of a continuous scale and a completely phased genotype for prediction of CYP2D6 enzyme activity and could potentially enable more accurate prediction of individual drug response.

Introduction

In personalised medicine, pharmacogenomics (PGx) is a crucial component that ensures the safety and efficacy of drug treatments based on patient's genetic profile [1,2]. The cytochrome P450 (CYP) isoenzyme 2D6, encoded by the polymorphic *CYP2D6* [3], is involved in the metabolism of 25–30% of commonly prescribed drugs [4]. Genetic variants in the *CYP2D6* gene, such as SNVs (single nucleotide variants), CNVs (copy-number variants) and structural rearrangements [3,5,6], may lead to differential activity of the enzyme Cytochrome P450 (CYP) 2D6 and thereby to altered drug response [7,8].

To translate *CYP2D6* variants into clinically actionable guidelines, they are assigned to standard haplotypes and predicted phenotypes. Haplotype assignment is performed based on *-allele nomenclature, catalogued by the Pharmacogene Variation Consortium (PharmVar), where each *-allele describes a predefined combination of variants [9,10]. Subsequently, the gene activity score (GAS) system assigns a score to each allele, with 0 for no activity, 0.5 for decreased, 1 for normal and 2 for increased activity [11]. Predicted phenotypes are assigned based on the combination of the two inferred allele activities and are summarized into 4 different CYP2D6 metabolizer categories [10,12]: poor metabolizer (PM), intermediate metabolizer (IM), normal metabolizer (NM) and ultra-rapid metabolizer (UM). However, 6 to 22-fold unexplained intra-category variability in enzyme activity and considerable overlap in activity between phenotypes remains [13]. Moreover, a recent study of twins has shown that although 91% of CYP2D6 metabolism is hereditary, GAS-based inferred phenotypes only explain 39% of variability in CYP2D6 enzyme activity [14]. Similar trends in missing heritability have been shown for other genes involved in CYP450-mediated drug metabolism [15,16]. This is partly due to rare genetic variants that are not catalogued in the current *-allele nomenclature [17]. A major limitation of the current methodology is the loss of a considerable amount of information in the categorization. Therefore, as has been suggested previously, a continuous phenotype prediction rather than a categorical model is likely to improve the prediction of CYP2D6 enzyme activity [8,18]. A convolutional neural network is highly suitable for this type of phenotype prediction from genetic data [19,20]. Although previous approaches of deep learning in pharmacogenomics were aimed at automated *-allele assignment or to predicting the residual activity of conventional *-alleles [21,22], we propose a strategy to predict CYP2D6 enzyme function on a continuous scale using full gene sequencing data and a neural network, omitting *-alleles completely.

Results

Conventional categorical phenotype predictions

To study the explained variability of conventional phenotype assignment, we included 561 subjects of European ancestry from the prospective CYPTAM study, which investigated the relationship between *CYP2D6* genotype and outcome of breast cancer treatment with tamoxifen (Supplementary Figure S5.1) [23]. The metabolism of tamoxifen is complex and involves multiple other enzymes (Supplementary Figure S5.2). Nonetheless, the conversion from desmetyltamoxifen to endoxifen is dominated by CYP2D6 [24,25]. Therefore we inferred CYP2D6 enzyme activity by using the ratio between the metabolites endoxifen and desmethyltamoxifen (Metabolic ratio (MR)) [26]. The influence of common variants in other enzymes (CYP3A4, CYP3A5, CYP2C19, and SULT1A1) has been investigated but did not have a profound effect on this conversion [27,28].

To fully resolve the *CYP2D6* paternal and maternal alleles, we applied long-read sequencing using germline DNA [5], which yielded comparable predicted phenotype results to orthogonal testing (Kappa-coefficient: 0.95 $p=3.5x10^{-169}$) (Supplementary Figure S5.3 and Supplementary Table S5.1). Classification of patients into conventional metabolizer categories resulted in 54.4% ($R^2=0.54$) explained variability in CYP2D6 enzyme activity (Figure 5.1A, Table 5.1).

Table 5.1: Regression results explaining CYP2D6 enzyme activity using different methods
All predicted phenotypes are based on PacBio long-read sequencing data. Consensus guidelines were used to assign the conventional four phenotype categories and gene activity scores. A neural network trained on CYPTAM data was used for the prediction of a continuous phenotype. Per-allele contributions from the neural network were added together (similar to the gene activity score) to predict the effect of a continuous gene activity score per allele using an additive model. For CYPTAM-BRUT, patients with unknown inhibitor use were excluded from this analysis (n=16). All values are the adjusted R^2 based on linear regression, with a p-value cut-off of 0.05 for significance.

	CYPTAM (n=561)	CYPTAM-BRUT no inhibitors (n=127)	CYPTAM-BRUT inhibitors (n=24)	Venlafaxine (n=69)
Categorical phenotype	0.5443, $p=2.78x10^{-95}$	0.3483, $p=4.517x10^{-12}$	0.07752, p=0.1009	0.5461, $p=8.09x10^{-12}$
Conventional gene activity scores	0.6659, $p=1.07x10^{-127}$	0.492, $p=2.52x10^{-20}$	0.1583, p=0.0308	0.6354, $p=1.54x10^{-16}$
Continuous phenotyping prediction	**0.7885, $p=6.14x10^{-191}$**	**0.6618, $p=1.99x10^{-31}$**	0.1633, p=0.0286	**0.6385, $p=1.15x10^{-16}$**
Gene activity scores predicted (predicted allele activities in an additive model)	0.7278, $p=2.59x10^{-160}$	0.6012, $p=6.15x10^{-27}$	**0.1929, p=0.0183**	0.6064, $p=2.028x10^{-15}$

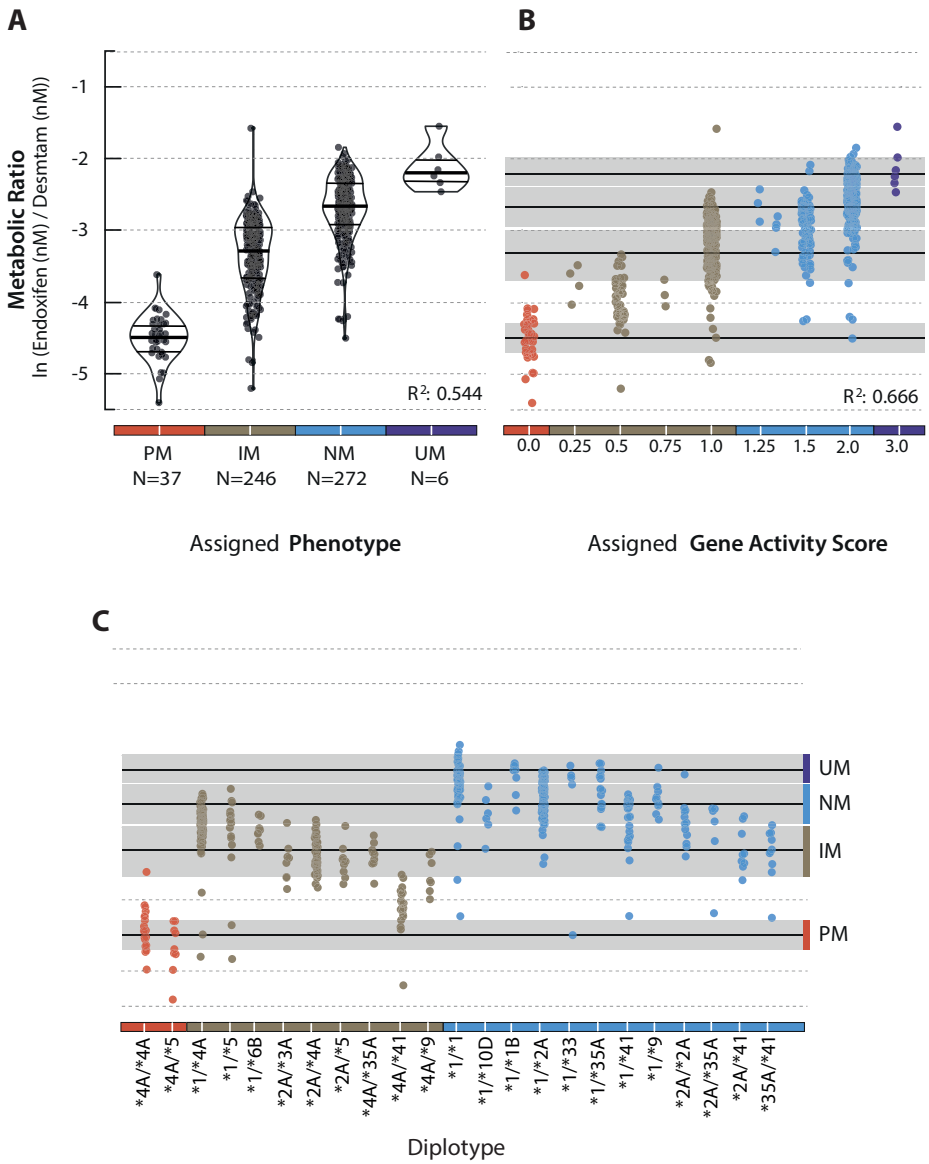


Figure 5.1: CYP2D6 activity based on conventional CYP2D6 metabolizer categories, gene-activity scores, and diplotypes

Explained variability of CYP2D6 activity in the CYPTAM-cohort, based on **(A)** conventional phenotype categories and **(B)** gene activity scores (n=561). **(C)** Tange in enzyme activity within common (>5% occurrence) diplotypes. Metabolic ratio (ln(Endoxifen (nM)/Desmethyltamoxifen (nM)) serves as proxy for CYP2D6 enzyme activity. Gene activity scores and phenotype predictions are based on *-allele nomenclature and Dutch Pharmacogenetic Working Group translations using PacBio long-read sequencing data. R^2: R^2adjusted based on linear regression. **(A)**: Violin plots display observation density, lines represent the median and inter quartile range. **(B and C)**: black lines represent median, grey area represents 95% confidence interval. PM: Poor Metabolizer, IM: intermediate metabolizer, NM: normal metabolizer, UM: ultra-rapid metabolizer.

Although the GAS system performed better than the 4 metabolizer categories (R^2=0.66), a considerable amount of variability in enzyme activity within each predicted phenotype category remained unexplained (Figure 5.1B, Table 5.1). Stratifying the phenotype categories into diplotypes showed that CYP2D6 activity varied substantially within identical diplotypes (Figure 5.1C). This suggested that a large proportion of the variability in enzyme activity within metabolizer phenotypes is already introduced when assigning haplotypes, with individuals carrying the same diplotype displaying phenotypes ranging from normal metabolizers to poor metabolizers.

A continuous scale improves phenotype predictions

To increase the explained variability in CYP2D6 enzyme activity, we developed and trained a neural network consisting of two parts (Supplementary Figure S5.4). The first part assigns contribution scores to individual alleles and variants and the second part combines paternal and maternal allelic scores into a predicted MR. Both parts were trained simultaneously on data generated from the CYPTAM-cohort. By including all observed variants independent of predefined haplotypes, the explained variability increased to 79% (R^2-adjusted = 0.79 (Figure 5.2A, Table 5.1)). Inter-individual variability is reflected by the range of observed MR in individuals with the same genetic make-up based on the 77 variants which are considered by the neural network (equal predicted MR). Additionally, there was large overlap in the predicted MR between individuals from the conventional IM and NM categories. The error rate (|observed MR – predicted MR|) was consistent over the range of the measured phenotype, with the exception of several subjects (n=16 (2.9%) outside of confidence interval) with a lower observed CYP2D6 activity than predicted (Figure 5.2A).

Allele contribution scores generated in the first part of the model were scaled to be comparable to the conventional GAS system (ranging 0–2). Allele contribution scores predicted by the model showed a deviation from the conventional GAS assignments for multiple *-alleles (Figure 5.2B).

For example, the *2A allele has a conventional GAS of 1.0 representing a fully active allele. However, the predicted allele contributions ranged from 0.60 to 0.90, accounting for variants which are not included in the reference *2A haplotype. Similarly, the predicted average contribution for *41 is 0.34 (95% CI (confidence interval): 0.33–0.36), whereas the conventional assignment for the *41 allele is 0.5 [9]. The same holds for the relatively rare *59 allele, currently regarded as decreased activity assuming a GAS of 0.5 whereas we predicted the activity to be 0.20 (95% CI: 0.19–0.22). The use of allele contribution scores on a continuous scale in an additive model improved the prediction of enzyme activity to 73% (Figure 5.2C). There are studies underway that aim to investigate the effect of

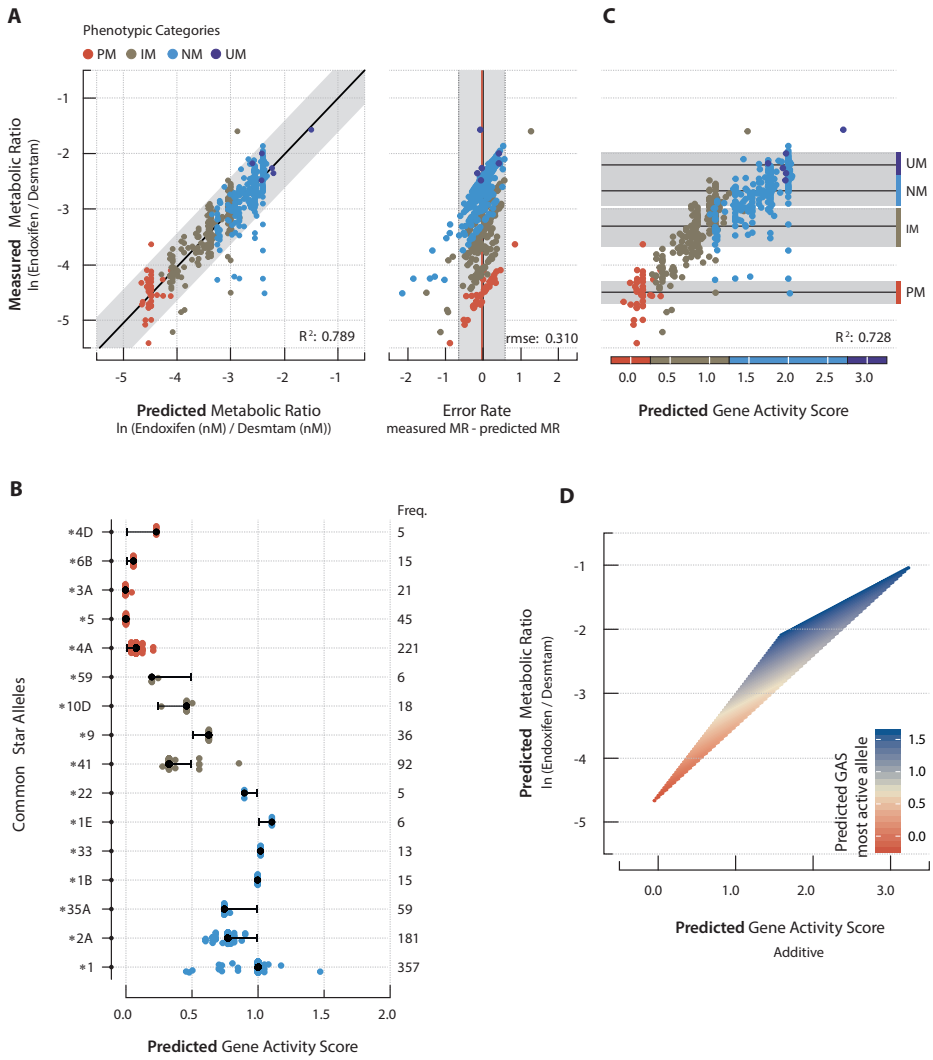


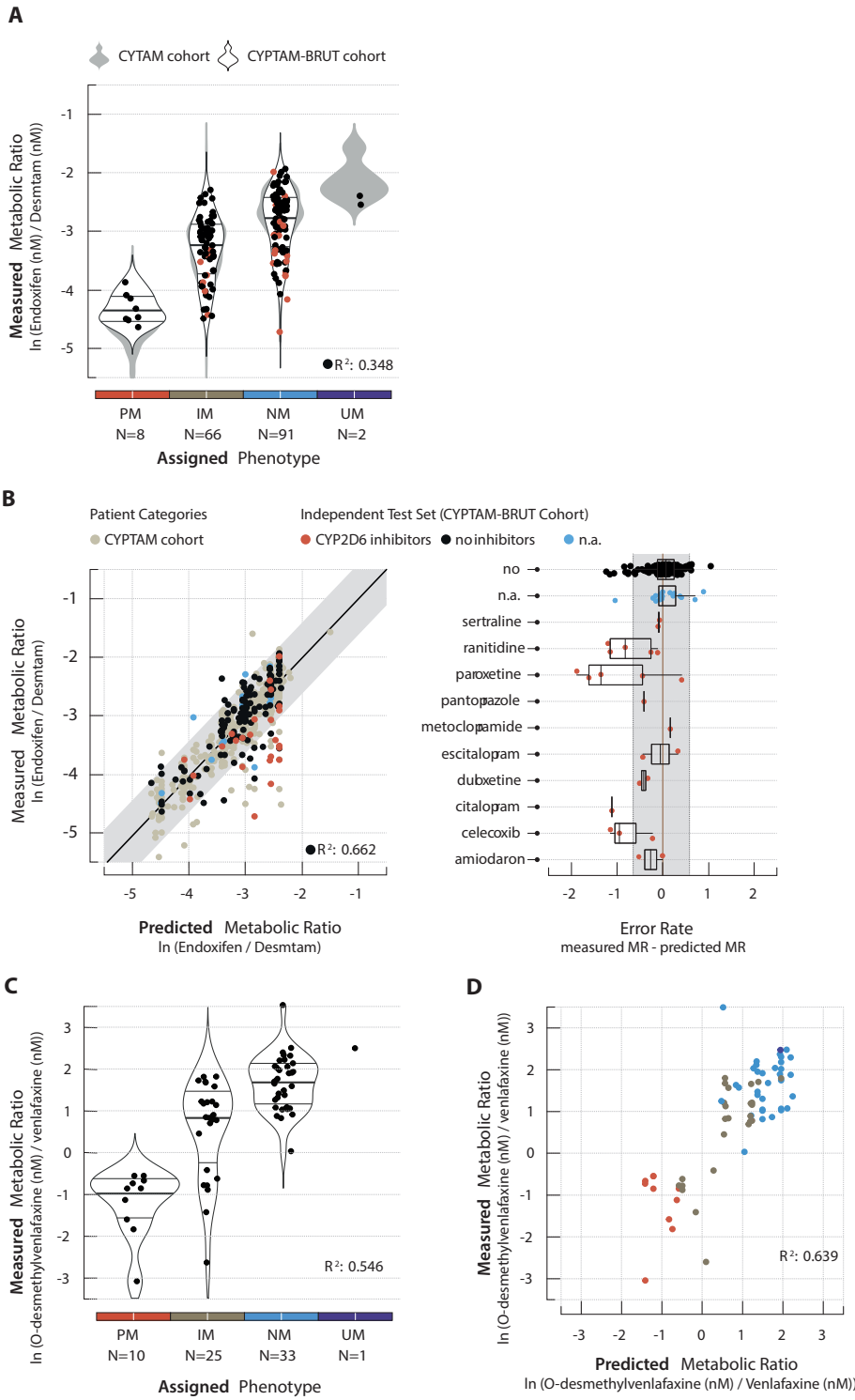
Figure 5.2: Neural network predictions for the CYPTAM cohort
(A) The model predicts the metabolic ratio (ln(Endoxifen (nM)/Desmethyltamoxifen (nM)) as a proxy for CYP2D6 enzyme activity on a continuous scale, with a consistent error rate over the entire range (n=561). Colouring is based on conventional *-allele assignments and metabolizer categories and shows a continuous transition from one category to the next, with overlap between the IM and NM groups. **(B)** Explained variability in enzyme activity using an additive model for the predicted allele contributions (n=561), such that predicted gene activity score = predicted contribution allele 1 + predicted contribution allele 2. **(C)** Predicted contributions per allele grouped in conventional *-allele assignments. **(D)** Comparison of the neural network predicted gene activity score in an additive model with the neural network predicted metabolic ratio. Where the predicted gene activity score additive = predicted contribution allele 1 + predicted contribution allele 2, the predicted metabolic ratio is the final outcome of the neural network and the colours represent the activity of the most active allele. R^2: adjusted R^2 based on linear regression. Rmse: root mean square deviation. **(A)** and **(B)**: blacklines represent median, grey area represents 95% confidence interval MR: Metabolic Ratio (of ln(Endoxifen(nM)/ desmethyltamoxifen(nM))), PM: poor metabolizer, IM: intermediate metabolizer, NM: normal metabolizer, UM: ultra-rapid metabolizer.

specific alleles on a continuous scale [29]. However, simply applying an additive model to individual allele contribution scores may be an oversimplification of human physiology. The second part of the neural network can accommodate non-additive combinations and therefore identify more complex relations between 2 alleles. Indeed, when the sum of allele contributions remains the same, a higher overall activity was observed when one of the alleles was fully active and one was fully inactive compared to two alleles with decreased activity (Figure 5.2D), which is in concordance with previous reports on IM phenotype variability [30].

Increased explained variability in independent samples and CYP2D6 substrates

To validate the model, 167 subjects of European ancestry receiving tamoxifen who participated in the CYPTAM-BRUT study [31] were sequenced using long reads and analysed with our neural network (Supplementary Figure S5.1 and Supplementary Table S5.1). In this cohort, patients were divided into two groups based on the use of CYP2D6 inhibitors that could influence the measured metabolic ratio of tamoxifen. Conventional phenotype predictions explained only 34.8% (R^2-adjusted = 0.35 of the variability in CYP2D6 enzyme activity (Figure 5.3A and Table 5.1) in subjects without concomitant CYP2D6-inhibiting drugs (n=127). Neural network-based phenotype prediction on a continuous scale resulted in an almost doubling of the explained variability (R^2-adjusted = 0.66) (Figure 5.3B and Table 5.1). These numbers were lower than those in the CYPTAM-cohort, most likely due to a lower sample size and lower density of observations in the extremes of the enzyme activity (PM and UM). For subjects using concomitant CYP2D6 inhibitors (n=24), 7.8% and 16.3% of the CYP2D6 activity could be explained by conventional and continuous phenotype prediction, respectively (Table 5.1). Moreover, there was substantial overlap between subjects with a negative deviation from the predicted enzyme activity in the CYPTAM-cohort and patients from the CYPTAM-BRUT cohort receiving concomitant treatment with a CYP2D6 inhibitor. This observation suggests that the concomitant use

Figure 5.3: Conventional and continuous predictions in replication cohorts
The explained variability of CYP2D6 enzyme activity for CYPTAM-BRUT (tamoxifen metabolism, n=167) based on **(A)** conventional phenotype categories and **(B)** on a neural network trained with data from the CYPTAM cohort. The influence of CYP2D6 inhibiting drugs on the overall enzyme activity shows overlap with CYPTAM samples with a negative deviation from the predicted enzyme activity. Error rates per CYP2D6 inhibiting drug give an indication of inhibitor potency (B, n=24 total). The explained variability of CYP2D6 enzyme activity for venlafaxine cohort (venlafaxine metabolism, n=69) based on **(C)** conventional phenotype categories and **(D)** on a neural network trained with data from the CYPTAM cohort. R^2: R^2-adjusted based on linear regression. In **(B)**: black lines represent median, grey area represents 95% confidence interval. In **(A)** and **(C)** Violin plots display observation density, lines represent the median and inter quartile range. PM: poor metabolizer, IM: intermediate metabolizer, NM: normal metabolizer, UM: ultra-rapid metabolizer.


A
CYTAM cohort
CYPTAM-BRUT cohort
Measured Metabolic Ratio
ln (Endoxifen (nM) / Desmtam (nM))
R²: 0.348
PM N=8
IM N=66
NM N=91
UM N=2
Assigned Phenotype
B
Patient Categories
CYPTAM cohort
Independent Test Set (CYPTAM-BRUT Cohort)
CYP2D6 inhibitors
no inhibitors
n.a.
Measured Metabolic Ratio
ln (Endoxifen / Desmtam)
R²: 0.662
Predicted Metabolic Ratio
ln (Endoxifen / Desmtam)
no
n.a.
sertraline
ranitidine
paroxetine
pantoprazole
metoclopamide
escitalopram
duloxetine
citalopram
celecoxib
amiodaron
Error Rate
measured MR - predicted MR
C
Measured Metabolic Ratio
ln (O-desmethylvenlafaxine (nM) / venlafaxine (nM))
R²: 0.546
PM N=10
IM N=25
NM N=33
UM N=1
Assigned Phenotype
D
Measured Metabolic Ratio
ln (O-desmethylvenlafaxine (nM) / venlafaxine (nM))
R²: 0.639
Predicted Metabolic Ratio
ln (O-desmethylvenlafaxine (nM) / Venlafaxine (nM))

of CYP2D6 inhibitors may explain the overestimated enzyme activity for several subjects in the CYPTAM-cohort. Stratifying the error between the predicted and observed enzyme activity per CYP2D6-inhibiting drug provides an estimate of the potency of the inhibitor (Figure 5.3B).

CYP2D6 enzyme activity is substrate specific and the effect on metabolism of a given variant varies per drug [32,33]. To assess substrate specificity of the neural network, we tested the performance on patients treated with a different CYP2D6 specific substrate, the antidepressant venlafaxine [34] (Supplementary Figure S5.1, Supplementary Table S5.1 and Table 5.1). In venlafaxine-treated patients of European ancestry, the explained variability of CYP2D6 activity increased from 54.6% (R^2-adjusted = 0.55) for conventional phenotype prediction to 63.9% (R^2-adjusted = 0.64) for phenotype predictions on a continuous scale (Figure 5.3C and Figure 5.3D). Although the explained variability improves with the continuous prediction compared to conventional categorization, the increase was limited.

In vitro validation of predicted variant contributions

We also queried the trained neural network to assess the contribution of individual variants to overall enzyme activity (Supplementary Data File S5.1). These contributions showed a wide range of effects in a pattern indicative of a continuous scale as opposed to an on/off effect (Figure 5.4A). To confirm the contributions predicted by the model, 4 variants and the *2 allele were expressed in HEK293 (human embryonic kidney) cells, which are completely devoid of any other metabolic enzymes [35,36], and incubated with bufuralol (Supplementary Table S5.2). The direction of our predictions (decrease or increase) was in concordance with the in vitro results, with accuracy increasing as the variant frequency increases. The predicted activity for *2 closely matched the observed activity in HEK293 cells (Figure 5.4B). For *CYP2D6**2 it is known that the normal activity of the allele is generally caused by the presence of an enhancer mutation causing an increase in expression which is almost in full linkage disequilibrium with the *2A allele [37,38]. This enhancer mutation is generally not included in in vitro experimental setups, leading to ower activity compared to wildtype. The presence of the enhancer mutation was not included in our neural network model, but was present in the population [39]. Nonetheless, in this study we observed decreased *CYP2D6**2A activity in vivo, suggesting that *2A activity might be substrate-dependent. This decreased activity for *CYP2D6**2A in tamoxifen metabolism has been observed previously [33].

For the gain of function variant Phe120Ile, the difference between the in vitro experiment and the neural network prediction was 8-fold, which might be explained by both the low allele frequency in the training cohort (n=3) and the substrate specificity of

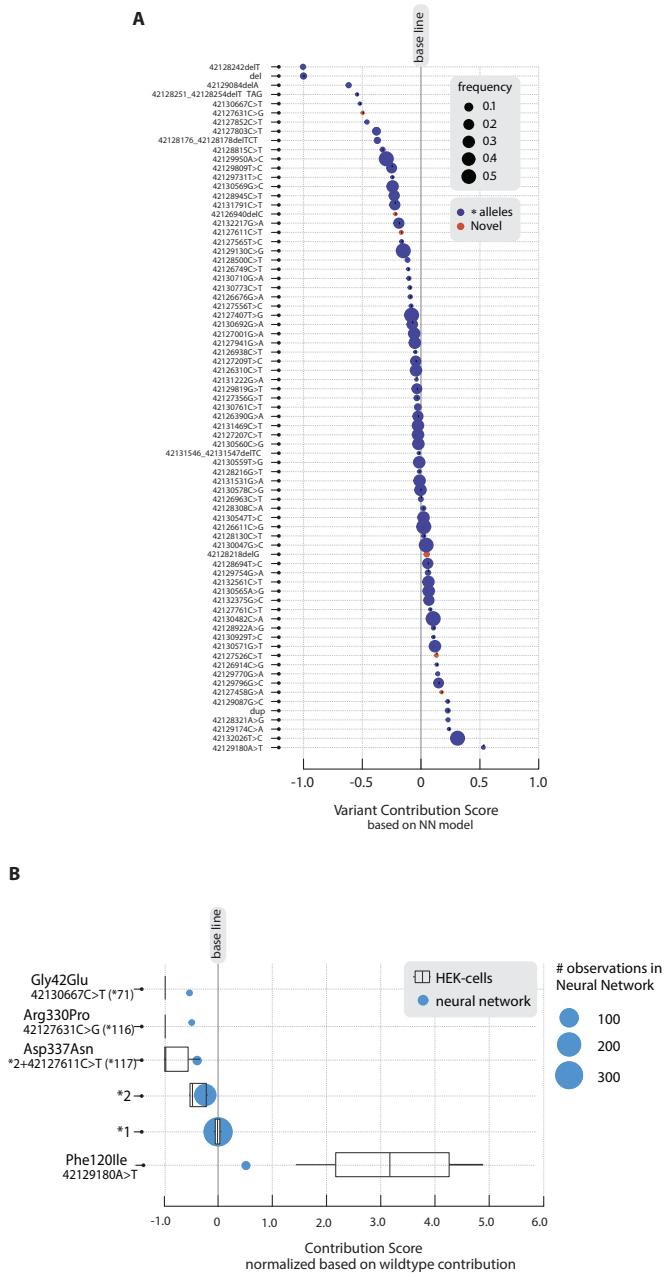


Figure 5.4: Contributions of individual variants

(A) Predicted contributions per variant included in the training of the neural network, with the absence of the variant set to 0.0 and a gene deletion set to -1.0 (n=78 variants). **(B)** In vitro validation of variants in HEK cells using bufuralol metabolism. Rate of bufuralol metabolism was normalized using the metabolic rate of the cells transfected with CYP2D6*1 cDNA as reference point of 1.0. Similar to the neural network contributions results were further scaled to have the absence of a variant (normal activity) set to 0.0 and the full absence of metabolism to -1.0. Incubations were performed in quadruplicate.

this variant. As the Phe120Ile variant is known to affect CYP2D6 substrates differently, this difference might be explained by the fact that the model was trained on tamoxifen metabolism whereas the in vitro experiment was performed using bufuralol as a proxy for enzyme activity [33,40,41]. The Phe120Ile amino substitution has previously been seen in *CYP2D6*49* where the allele has lower activity due to the presence of the Pro34Ser mutation and in *CYP2D6*53* where the Phe120Ile mutation is accompanied by an Ala122Ser [9]. These alleles were not present in our study population, as none of the subjects carried the other mutations in combination with the Phe120Ile mutation [33]. These results indicate that the model was able to detect the biological effect of variants with improved accuracy as the number of observations increased.

Discussion

The role of Cytochrome P450 enzymes in the metabolism of commonly prescribed drugs has provided a basis for developing a series of molecular and genotyping assays to predict patient drug responses. To date, genotyping and phenotype predictions for clinical PGx utilize a categorical approach where predefined genomic variants are screened to infer patient haplotypes according to the *-nomenclature and thereby the associated phenotype category [10]. However, in current clinical PGx, rare variants (minor allele frequency <1%) are often excluded and haplotype phasing is inferred rather than resolved. Our proof-of-concept study, based on allele-specific genotyping and continuous phenotype predictions powered by a neural network model that predicts tamoxifen metabolism, showed improvements in decreasing the missing heritability in CYP2D6-mediated metabolism. In the CYPTAM-cohort, we observed a 25% increase in explained variability compared to the conventional approach. Notably, our model achieved improvements of 31% and 11% in independent tamoxifen (CYPTAM-BRUT) and venlafaxine replication cohorts respectively in predicting patient drug metabolic rates and thereby potentially the drug response. Our model is agnostic to the sequencing technology used and could be applied to data obtained with alternative genotyping methods. However, it should be kept in mind that haplotype phasing might be less reliable with other sequencing methods. Moreover, problems might arise with the use of short-read sequencing due to the very similar structures of the flanking pseudogenes *CYP2D7* and *CYP2D8* as well as deletion and duplication variants of these genes.

The inhibition of tamoxifen metabolism by the use of concomitant CYP2D6 inhibitors could be estimated in the CYPTAM-BRUT cohort. Our results confirm that paroxetine has the strongest impact on enzyme function [42]. Guidelines on GAS adjustments based

on inhibitor use are available for CYP2D6 [43]. However, these GAS adjustments are not yet standard clinical practice. Moreover, inhibitor use might change over time whereas genetic make-up will not, making it of importance to be able predict enzyme activity based solely on genetics. In our study, subjects receiving concomitant treatment with a CYP2D6 inhibitor were outliers in the enzyme activity prediction graphs and the strength of each inhibitor matched the error between observed and predicted enzyme activity. These findings indicate that the model was learning the role of genetic information and it is expected that the contribution of other potential confounding variables to the overall model is minimal. In drug response management, it is critical to detect the role of concomitants in drug safety and efficacy. Thus, our results pave the way for reliable deconvolution of the expected drug metabolic rate and the impact of concomitants to better inform treatment strategy. Although the increase in explained variability in the CYPTAM-BRUT cohort was similar to that in the CYPTAM cohort, the R^2 of both the conventional model and the continuous model was lower in the CYPTAM-BRUT cohort compared to the CYPTAM cohort. This difference can be explained by the sample size as well as the density of patients in the extremes of the metabolic groups (PM and UM), leading to a higher variability and lower representation of the entire metabolic categories.

In the venlafaxine cohort, the gain in explained variability was limited (55% to 64%) compared to the other two cohorts. Both substrate specificity [32,33] as well as the limited sample size (n=69) in the venlafaxine cohort may contribute to this limited increase in explained variability. In addition, due to the limited sample size, the majority of samples were from intermediate and normal metabolizers, limiting the genetic diversity and thereby affecting the overall R^2. Moreover, venlafaxine is not solely metabolized by CYP2D6. Indeed, it has recently been shown that CYP2C19 has a significant contribution to the venlafaxine metabolism ($p<0.001$ for the majority of metabolizer groups) [44]. In this study we used the metabolic ratio between endoxifen and desmethyltamoxifen as a proxy for enzyme activity. Although the metabolic conversion of endoxifen to desmethyltamoxifen is CYP2D6-specific, other enzymes are involved in the metabolism of tamoxifen, thereby influencing the concentration of the metabolites used in our model as well [23,24,27,28]. Moreover, substrate specific effects might result in limitations regarding the generalizability of the model for all CYP2D6 substrates. Therefore, this ratio might not reflect the activity of CYP2D6 for other substrates. Nonetheless, the model still outperformed the current clinical standard as an increase in explained variability was also observed for the CYP2D6 substrate venlafaxine. To explore the full capability of the proposed neural network approach, large cohorts with CYP2D6-specific substrates are needed. In future studies, collecting real-world evidence from larger cohorts can facilitate our understanding of substrate specificity and the role of CYP2D6 in the metabolism of wider range of drugs.

Markedly, the contribution of individual alleles to the overall enzyme activity did not fully align with the additive model that is currently used for phenotype assignment. The model that adds the two individual allele contribution scores resulted in a lower explained variability than the combination of the two alleles in the neural network (R^2 of 0.73 and 0.79 respectively for the CYPTAM cohort). This finding can either be explained by a genetic component (for example up-regulation of *CYP2D6* expression or non-*CYP2D6* variants) or by a non-linear relation between CYP2D6 enzyme activity and the metabolic ratio. There is, however, no indication to assume a non-linear relation in CYP2D6 enzyme activity and tamoxifen metabolism [45,46]. Additionally, we observed variable activity between alleles which have been assigned the same *-haplotype. This interindividual variability in allele activity can be associated with exclusion of variants outside of the best matching *-haplotype. The overall activity of star-alleles deviated from defined categorical assignments and was suggestive of a more continuous nature. Rare and not yet catalogued variants might be responsible for this deviation as they are not included in *-haplotype definitions [17].

Our data show that, for a subset of variants and alleles, the predicted variant and allele contributions are in line with those in in vitro experiments. Nonetheless, the predicted activity based on the neural network and the in vitro data did not always align completely. A potential explanation for this is the role of substrate specificity. Different CYP2D6 substrates are affected by CYP2D6 variants to a different degree. For example, the impact of variants affecting the protein binding site might also be dependent on the size of the drug that is being metabolised, if the size of the binding site is decreased, smaller molecules might be less affected that larger molecules [47]. Another potential explanation may be that both in vitro predictions and neural network predictions are not perfect reflections of an in vivo situation.

It has been shown that in-vitro results can give a reliable indication of the in vivo effect of a specific variant [35,36]. Both the neural network predictions, which are in silico predictions based on in vivo data, and the in vitro results showed the same impact (decreased, normal, or increased activity) of the investigated variants. These results indicate that the neural network-based prediction can be valuable in evaluating the potential deleteriousness of SNPs in a similar way as in vitro experiments.

Multiple tools have been developed to predict PGx phenotypes from sequencing data, of which stargazer and Hubble are the most prominent [21,48,49]. Stargazer aims at assigning star-haplotypes and thereby phenotype categories based on catalogued genetic markers whereas Hubble has been trained on in vitro *-allele activity to predict phenotypes on a continuous scale. Our approach completely omits *-alleles and predicts in vivo phenotypes on a continuous scale, making it a unifying model from sequence to phenotype.

This study focused on the role of genetic variants within *CYP2D6* locus on variable drug response for tamoxifen and venlafaxine. However, besides *CYP2D6* genetics, additional genetic factors can also regulate CYP2D6 expression and thereby influence response to these drugs [50-52]. The eQTLs (expression quantitative trait loci) described by the GTEx (genotype-tissue expression) project could be a valuable resource to assess the impact of various genetic variants in gene expression in different tissues [53]. Moreover, recent studies have shown that several transcription factors (such as *TSPYLs* and *HNF4α*) can influence the expression of CYP enzymes [51,54]. However, limited samples of liver tissue in the GTEx database, where known PGx genes including *CYP2D6* are predominantly expressed, did not allow us to investigate the potential impact of variants found in our study. Notably, additional metabolic enzymes might also be involved in the drug metabolism. For tamoxifen, *CYP2C19*, *CYP3A4*, *CYP3A5* and *SULT1A1* can also play a role in several steps of the metabolic pathway. However, previous studies suggest that the role of these enzymes in tamoxifen metabolism may be limited [27,28,55]. Further studies on larger cohorts, where data on observed drug phenotypes as well as the genetic makeup of all responsible enzymes and potential transcription factors is available, may shed light on the role of other genetic variants in the metabolism of tamoxifen and venlafaxine. Last, the approach in this study is focussed on the PGx of the pharmacokinetic marker *CYP2D6*. Nonetheless, PGx of pharmacodynamic targets can also play a role in variability in drug response [2,56,57]. This field is currently less well understood due to the difficulties associated with phenotype definitions. For pharmacokinetics, this is simpler as the effect can be measured by drug concentrations. Nonetheless, potential impact of pharmacodynamics in the field of PGx should not be neglected and warrants further research.

Our study suggests that, for the gene and drugs included, the proposed strategy is capable of improving patient phenotype predictions by using a continuous scale for this prediction as well as offering insights on genetic variants the underlie variable drug response. However, several limitations do exist. First, this study focused on the role of variants within the *CYP2D6* locus and did not account for variants outside of this locus that can potentially influence variability in drug response, as discussed above. Moreover, drug response is not only influenced by genetic factors. Multiple additional factors such as substrate specificity, lifestyle, concomitant drug use, comorbidities, and epigenetics are known to play a role in patient treatment response [16,27,28,58,59]. Unfortunately, our study was limited in regards to the ability to include these factors due to the sample size and lack of data on co-medication and comorbidities. Additionally, our sample size precluded characterization of all rare variants. Last, the frequency of (pharmaco)genetic variants is known to vary between different ethnicities [60,61]. However, our study included only individuals of European ancestry and therefore did not include any variants which might be specific

to other ethnicities. In the future, larger cohorts of varied ethnicity with more extensive clinical data as well as a more comprehensive genetic makeup of the pharmacogenes could improve drug response prediction models even further, although it is unlikely that all environmental factors can be accounted for. Last, our study was a proof-of-concept for CYP2D6 and tamoxifen and venlafaxine, studies applying the same approach to other gene-drug combinations should be conducted to confirm the value of a neural network-based approach for PGx.

Materials and methods

Study design

The aim of this study was to develop a model to predict CYP2D6 enzyme activity on a continuous scale and compare this approach to the conventional categorical methods. Genetic markers in the *CYP2D6* locus were used as the predictors. For the outcome, the metabolic ratio of CYP2D6 substrates was used as a proxy for enzyme activity. Existing cohorts were included based on availability, therefore sample size calculations were not performed. A cohort of 608 individuals was used for the development of the model, and two independent cohorts of 225 and 78 individuals were used for replication. The CYPTAM protocol was approved by the Institutional Review board of the Leiden University Medical Center (LUMC). The CYPTAM-BRUT protocol was approved by the Institutional Review board of the Leuven University medical center. Venlafaxine samples were collected in routine clinical care at Catharina Hospital, Eindhoven, the Netherlands. The medical ethics committee of the Catharina Hospital provided a waiver for consent as samples and data for study purposes were already available, according to the code of conduct for responsible use of human tissue and medical research (fedora.org).

Study cohorts

The data used in this study originated from one main cohort and two independent cohorts of European ancestry. The main study cohort, the CYPTAM-cohort, consisted of 608 subjects for whom DNA material was available (The Netherlands National Trial Register: NTR1509) [23]. In short, the multicenter prospective CYPTAM study recruited subjects receiving tamoxifen as an adjuvant breast cancer therapy to investigate the association between *CYP2D6* genotype, endoxifen serum concentration and clinical outcomes. The first replication cohort, the CYPTAM-BRUT cohort, consisted of 225 subjects recruited in a study investigating the association between *CYP2D6* genotype and endoxifen serum

concentration on response rate to tamoxifen in postmenopausal women (Clinicaltrials.gov: NCT00965939) [31]. The second replication cohort, the venlafaxine cohort, consisted of 78 Dutch subjects taking venlafaxine. Samples were collected as part of routine patient care at the Catharina Hospital, Eindhoven, the Netherlands. DNA samples and accompanying data were de-identified before transfer to the LUMC for analysis.

Drug metabolite measurements

For both the CYPTAM and CYPTAM-BRUT cohort, steady state through concentrations of tamoxifen and metabolites were measured with a validated high performance liquid chromatography-tandem mass spectrometry upon study inclusion. All measurements were performed at the LUMC department of Clinical Pharmacy and Toxicology. In total 4 compounds were measured: tamoxifen, 4-hydroxytamoxifen, O-desmethyltamoxifen and endoxifen. A total of 0.2 ml of each serum sample was mixed with 0.5 ml of 0.1 M $ZnSO_4$ and 0.2 ml of the internal standard working solution 4-D5-IS. After mixing for 3 min on a vortex mixer, the mix was centrifuged at 13,000 rpm for 5 min at room temperature. A volume of 20 µl supernatant was injected into the HPLC instrument. Chromatographic analysis was performed using a Waters Micromass Quattro micro API Tandem MS equipped with a Dionex P680A DGP-6HPLC pump, Dionex Ultimate 3000 autosampler and a Diones Thermostated Column Compartment. Separation of the analytes from potentially interfering serum components was achieved using a Waters X-bridge Column (3.5 µm, 4.6 x 50 mm) with a Spark HySphere C18 HD pre-column (7 µm) in a Phenomenex holder. The mobile phase consisted of 25% solution A (0.1% formic acid + 2 mM ammonium acetate in H_2O) and 75% of solution B (0.1% formic acid + 2 mM ammonium acetate in methanol) and was delivered at a flow rate of 0.4 ml/min. Concentrations were normalized to nM and metabolic ratios calculated to reflect the rate of conversion from one metabolite to the next.

For the venlafaxine cohort, plasma concentrations of venlafaxine and its metabolite O-desmethylvenlafaxine were determined as part of routine clinical care. Concentrations were determined with a validated ultra-performance liquid chromatography-tandem mass spectrometry method. Clozapine-D4 dissolved in acetonitrile was used as internal standard in a concentration of 0.1 mg/L. To 100 µl of each plasma sample, a volume of 300 µl of internal standard solution was added and vortex-mixed for 30 seconds. After centrifugation for 10 min at 10,900 rpm, a volume of 200 ul of the supernatant was mixed with 200 µl of a 5 mM ammoniumacetate solution and 10 µl of this mix was injected on the UPLC-MS/MS. Chromatographic analysis was performed using a Waters Acquity UPLC with a BEH C18 (2.1 x 100 mm, 1.7 µm) column at 40°C. The mobile phased consisted of 90% solution A (5 mM ammoniumacetate + 0.05% formic acid) and 10% solution B (acetonitril 100%)

and was delivered at a flowrate of 0.35 ml/min. Concentrations were normalized to nM. All samples were analysed at the Catharina hospital department of Clinical Pharmacy.

DNA sample processing

Germline DNA isolation from blood was performed previously for the main studies and for routine clinical use. Remaining DNA samples were collected and transferred to the LUMC for sequencing. All samples were sequenced with Pacific Bioscience's (PacBio) SMRT-sequencing technique using full length *CYP2D6* amplicons [62]. PacBio sequencing enables the identification of all variants in the locus, including those in difficult and repetitive regions in addition to obtaining fully phased paternal and maternal alleles [5]. To obtain *CYP2D6* amplicons, three separate two-step PCR reactions were executed, one for full length amplicons and two for Copy Number Variants (CNV) using a similar protocol to Buermans et al. [5]. The current protocol differed in regards to the scale at which the analysis was performed which required larger sets of barcode primers. Additionally, the two replication cohorts were sequenced using the PacBio Sequel platform as opposed to the RSII platform which was used for the study by Buermans et al. and the training cohort. All primers used were based on previous research [63,64] and ordered from Integrated DNA Technologies (IDT) [65] (Supplementary Table S5.3).

The *CYP2D6* specific primers were designed to generate a 6.6 kB fragment covering the entire *CYP2D6* locus including upstream and downstream regions [63,64]. Target regions were amplified using the Takara LA Taq DNA polymerase kit [66]. A 10µl reaction volume contained 50–100 ng DNA, 1x PCR buffer with $MgCl_2$, 0.4mM dNTPs, 0.4 µM of both of the full length *CYP2D6* primers and 0.4 U Takara La taq. PCR cycle parameters were 3 min at 95°C, followed by 30 cycles of 10 sec 98°C and 15 min 68°C, finished with 15 min at 68°C. Subsequently, amplicon barcoding was performed using M13-tailed primers. These barcode primers were introduced in a second PCR with identical conditions to the first, using 1 ul of the first PCR product and 5 cycles of amplification.

CYP2D6 gene deletions were identified with a duplex PCR. The primer set consisted of *CYP2D6*-deletion specific primers and an internal control (IC) [63,64]. Target regions were amplified using the KAPA long range hotstart kit from kapa biosystems (REF: KK3502) [67]. The 10 µl reaction volume contained 50–100 ng DNA, 0.5x PCR buffer, 1.7 mM $MgCl_2$, 0.3 mM dNTPs, 0.5 uM of *CYP2D6*-deletion specific primers, 0.375 µM of IC primers and 0.025 U Kapa Hotstart polymerase. Cycle parameters were 3 min at 95°C, followed by 30 cycles of 15 sec 95°C and 10 min 68°C.

CYP2D6 gene duplication and *CYP2D6*/*CYP2D7* fusion gene conformations were identified using a triplex PCR protocol. The primer set contained the *CYP2D6* full length

primers, *CYP2D6* duplication primers and *CYP2D6*/*CYP2D7* fusion gene primers. The 10 µl reaction volume contained 50–100 ng DNA, 0.5x PCR buffer, 1.7 mM $MgCl_2$, 0.3 mM dNTPs, 0.5 µl DMSO, 0.5 µl of the *CYP2D6* full length forward primer and 0.75 µl of the reverse primer, 0.375 µl of both *CYP2D6*-duplication specific primers, 0.5 µl of the CYP2D6 fusion gene primer and 0.025 U Kapa Hotstart polymerase. PCR conditions were identical to the duplex PCR.

Presence of CNVs and fusion genes was assessed on a 0.7% agarose gel with ethidium bromide staining, set at 100 mV with a 55 min run time. CNV and fusion gene positive samples, identified as additional fragments besides a full length or IC fragment, were selected for the subsequent barcoding PCR. For the selected samples of both the duplex and triplex PCR, barcoding was done with M13-tailed primers. Identical conditions to the first PCR were used with 1 ul of PCR product from the first PCR and 5 cycles of amplification.

Barcoded amplicons were equimolar pooled into a full-length pool and a CNV and fusion genes pool. For the CYPTAM-cohort, one pool of full-length samples per 96-well plate was made and one pool for all CNVs and fusion genes. For CYPTAM-BRUT and the venlafaxine-cohort, one pool with all full-length samples and one pool for all CNV and fusion gene samples of both cohorts was made. All pools were concentrated with Ampure XP beads (Agencourt). For the full-length fragment, additional size-selection was performed using BluePippin (Sage Science) to remove all fragments shorter than 5kB prior to pooling with the CNV and fusion gene amplicons. SMRTbell library preparation was performed on 500 ng purified and size-selected PCR pool following the procedure & checklist – Amplicon template preparation and sequencing (PN 100-801-600 Version 04, Pacific Biosciences) and using SMRTbell template Prep Kit 1.0-SPv3 [62]. The final SMRT library was sequenced on the PacBio RSII for the CYPTAM-cohort and on the PacBio Sequel for the replication cohorts. For RSII, libraries were sequenced using sequencing primer V2 and P6-C4 chemistry with a movie time of 6hr, with a maximum of 96 samples per SMRT cell [62]. For Sequel, libraries were sequenced using sequencing primer V3, sequencing kit 3.0 and binding kit 3.0 on a 1M v3 LR SMRT cell with a movie time of 20 hr, with a maximum of 288 samples per SMRT cell [68]. Deletions, duplications and hybrids were analysed on a separate SMRT cell for all cohorts.

Data preprocessing

The full pipeline for downstream processing is available at DOI: 10.5281/zenodo.4787186. All downstream processing was run on a high-performance computing cluster running the sun grid engine. Raw sequences were demultiplexed using LIMA followed by the CCS

tool to generate CCS sequences. The subsequent haplotype phasing was performed using a custom pipeline which utilizes the CCS sequences to identify molecules originating from the same allele. Subreads of the CCS sequences were used to generate high quality phased allelic sequences per allele per individual using subreads of all molecules belonging to the same allele. Allelic sequences showing signs of disjoint sequences or chimeras were flagged. Per subject all phased allelic sequences were saved and plotted based on genomic distance.

Phased sequences were aligned to the *CYP2D6* sequence from GRCh38 and variants were called. A semi-global alignment was performed using biopython pairwise2, alignments were polished to ensure consistent indel positioning. Pharmacogenomic haplotype assignments were made based on PharmGKB translation tables [12]. For all haplotypes, the *-allele with a perfect match based on all variants observed was assigned, where the number of variants is decisive in the case of multiple perfect matches. When no perfect match is found the *1 haplotype was assigned. All identified variants were run through VEP (variant effect predictor) to determine their potential impact on protein function [69]. Variants were flagged as 'known' for variants in *-allele nomenclature, 'novel' for variants not in *-allele nomenclature, 'in polymer region' for variants located in homo-polymer regions.

The phased alleles were separated from chimeras and disjoint sequences by manual curation based on genomic distance plots and the presence of chimeras and disjoint flags. A cut-off of at least 10 molecules per allele and 10 passes per molecule was used to determine the reliability of the sequences. In the presence of gene deletion, the second allele was annotated as 'deletion'. A duplication, determined based on the number of molecules observed per allele, was annotated as 'duplicated'. Subjects identified as carrying a *CYP2D6*/*2D7* fusion gene were annotated as 'hybrid'. Selected alleles were linked to the clinical data based on subject specific barcodes, resulting in one datafile per cohort containing clinical data, selected alleles and haplotype calls.

Prediction models

For further analysis, samples were selected based on the presence of full length *CYP2D6* sequences, the absence of *CYP2D6*/*CYP2D7* conversions and fusion genes, and on the presences of clinical data regarding drug metabolism (n=561 for CYPTAM, n=167 for CYPTAM-BRUT, n=69 for venlafaxine). For each cohort the clinical datasets containing metabolite blood concentrations were merged with the sequencing data containing the assigned haplotypes.

Conventional method

For the CYPTAM-cohort, haplotype and phenotype assignments based on PacBio sequencing data were compared to calls from the Roche Amplichip which were determined previously. To assess explained variability based on conventional phenotyping, the same methods were applied to all three cohorts. For the CYPTAM-BRUT cohort data on concomitant use of CYP2D6 inhibiting drugs was available, based on which the cohort was split into 'non-inhibitor users', 'inhibitor users' and 'unknown inhibitor use'.

For all cohorts the same methods were applied. Haplotype calls were translated into Gene Activity Scores (GASs) and predicted phenotype categories based on the CPIC and DPWG consensus [8,70,71]. A GAS of 0.0 was assigned to non-active alleles, 0.5 to decreased activity, 1.0 to normal activity and 2.0 to increased activity alleles. Subsequently the scores per allele were combined into the overall GAS by adding them together, followed by a translation into phenotype categories. Based on the consensus paper of the Dutch Pharmacogenetics Working Group (DPWG) and the Clinical Pharmacogenetics Implementation Consortium) one of 4 clinically implemented phenotype categories was assigned: poor metabolizer (PM, GAS = 0.0), intermediate metabolizer (IM, GAS = 0.5–1.0), normal metabolizer (NM, GAS = 1.5–2.5) or ultra-rapid metabolizer (GAS = 3.0) [8].

As a proxy for CYP2D6 enzyme activity, the metabolic ratio of the most CYP2D6-specific conversion of either tamoxifen or venlafaxine metabolism was used. Although the metabolism of tamoxifen (desmetyltamoxifen to endoxifen) is dominated by CYP2D6, other enzymes play a minor role in the tamoxifen metabolism and therefore in the metabolite concentrations (Supplementary Figure S5.2) [27,28]. For the CYPTAM and the CYPTAM-BRUT cohorts, the log of the metabolic ratio of the conversion from desmethyltamoxifen to endoxifen ((ln(Endoxifen (nM)/ Desmethyltamoxifen (nM))) was used as a proxy for CYP2D6 enzyme activity [26,71]. Log transformation was performed to normalize the data. There are no indications to assume non-linear kinetics of endoxifen formation by CYP2D6 [45], in fact the kinetics of all other metabolites are linear [46]. Additionally, the metabolic ratio as used in this study was shown to stay consistent with dose increase for all phenotypes [30,72], making it a suitable proxy for enzyme activity. Last, it is expected that intra-individual variability of CYP2D6 enzyme activity is limited, making one measurement at steady state a suitable approach [73-75].

For the venlafaxine cohort, the log of the metabolic ratio for the conversion from venlafaxine to desmethylvenlafaxine (ln(O-desmethylvenlafaxine (nM) /venlafaxine (nM))) was used.

Neural network

From the selected alleles per individual, a dataset was generated indicating the presence (1) or absence (0) of every variant observed in the entire cohort (including deletions and duplications). From these variants a selection is made, variants were excluded if they adhere to the following rules: located in homopolymer regions or not in *-allele nomenclature and synonymous, intronic, located upstream or downstream. These were excluded to prevent confounding from irrelevant variants in the development of the neural network. Variants were included if they were part of the *-allele nomenclature or if they were additionally nonsynonymous, frameshifts or splice sites variants.

The neural network was build using Keras (https://github.com/keras-team/keras, version 2.2.4) with the TensorFlow (https://github.com/tensorflow/tensorflow, version 1.12.0) backend. It uses the selected variants (Supplementary Data File S5.1) per allele as predictors (n=78) and the measured metabolic ratio (ln(Endoxifen (nM)/Desmethyltamoxifen (nM))) as a surrogate for CYP2D6 enzyme activity and the outcome variable of the model. The model was comprised of 2 parts (Supplementary Figure S5.3). The first consisted of two interpreters, one per allele. These interpreters use all selected variants per allele as input data and combine them into an allele contribution. The second part was the combiner model which combined the two allele contributions to predict the metabolic ratio. The model was trained with the data from the CYPTAM-cohort and both parts were trained simultaneously. 10-fold cross validation with 100 cycles both with and without internal hold-out was performed and showed no signs of overfitting (Supplementary Figure S5.5). Shap (Shapley Additive explanation)-values were extracted and normalized to define allele contributions. 0.0 was assigned to a gene deletion and 1.0 to a fully wildtype allele. Variant contributions were normalized accordingly, resulting in the sum of variant contributions per allele corresponding to the allele contribution.

For both replication cohorts, the same variants as which were used during the training were included in the selection. For the venlafaxine cohort, the predicted metabolic ratio is translated with a linear transformation into the metabolic ratio for venlafaxine (ln (O-desmethylvenlafaxine (nM) /venlafaxine (nM))).

In vitro validation

To confirm the contribution of individual variants as predicted by the neural network, four high impact variants and the *2 allele were tested in vitro. Variants were selected based on the following criteria: the predicted contribution had to be ≥0.2 or ≤-0.2, without linkage disequilibrium with a known causal variant and potentially causal (for example missense or frameshift); both gain of function and loss of function variants were included. Variants

selected were: g.42130667C>T, g.4212761C>G, g.42127611C>T and g.42129180A>T as well as the *2 allele. Site mutagenesis was performed on pCMV4 *CYP2D6*1* plasmid [76] with QuikChange II Site-Directed Mutagenesis Kit (Agilent). Plasmid cDNA encoding variants with the following amino acid exchanges were created: Arg330Pro (g.42127631C>G), Gly42Glu (g.42130667C>T) and Phe120Ile (g.42129180A>T). The Asp337Asn exchange was performed using pCMV4 *CYP2D6*2* as template. Mutagenesis primers and selected variants are listed in Supplementary Table S5.4. Variants were expressed in HEK293 cells grown in DMEM 6046 (Sigma) containing 1 g glucose/l, 10% fetal bovine serum, and penicillin/streptomycin (100 IU/ml, 100 mg/ml) to a confluence of 60–70%. pCMV4 vectors containing the variants were transfected using Viromer Red (Lipokalyx) according to manufacturer's protocol. Cells were harvested after 24–48 hours incubation were stored at -80°C. Cell pellets were resuspended in 0.1 M sodium phosphate buffer followed by sonication for 20 x 1 sec and were centrifuged at 800 x g. Incubations were performed with 800 x g supernatant corresponding to 25–125 µg of protein, 0.1 M sodium phosphate buffer, 50 µM bufuralol (racemate), and 1 mM NADPH in a total volume of 150 µl. reactions were linear for at least 5 hours and were terminated by addition of 14 µl of 70% perchloric acid. After centrifuging the supernatant was analysed by high performance liquid chromatography as previously described [77]. The amount of CYP2D6 apoprotein of the different allelic variants were determined using sodium dodecyl sulfate polyacrylamide gel electrophoresis and Western blot analysis. Residual CYP2D6 activity was assessed and normalized with the average activity of the *-allele set at 1.0 to allow for comparison with the neural network predictions.

Statistical analysis

To compare amplichip and PacBio based haplotype calls, cohen's kappa was used, with a significance cut-off of $p<0.05$. The amount of explained variability in CYP2D6 enzyme activity for all phenotype predictions was assessed using linear regression, assuming a linear relation between predicted phenotypes and observed metabolic ratio. For the conventional approach, two different models were assessed, the first based on the clinical phenotype categories (PM, IM, NM, UM), the second based on overall GAS. For the neural network approach, the explained variability for all cohorts was assessed using linear regression with the predicted metabolic ratio as predictor and the observed metabolic ratio as outcome. Explained variability was expressed as R^2-adjusted, using a $p<0.05$ cutoff for significance. The error rate of the model was expressed as the rmse (root-mean-square error). All statistics were performed using R version 4.0.2. The haplotyping pipeline and neural network were developed using Python 2.

References

1. Relling, M.V.; Evans, W.E. Pharmacogenomics in the clinic. *Nature* **2015**, *526*, 343-350, doi:10.1038/nature15817.
2. Roden, D.M.; McLeod, H.L.; Relling, M.V.; Williams, M.S.; Mensah, G.A.; Peterson, J.F.; Van Driest, S.L. Pharmacogenomics. *Lancet (London, England)* **2019**, *394*, 521-532, doi:10.1016/s0140-6736(19)31276-0.
3. Gaedigk, A. Complexities of CYP2D6 gene analysis and interpretation. *Int Rev Psychiatry* **2013**, *25*, 534-553, doi:10.3109/09540261.2013.825581.
4. Ingelman-Sundberg, M. Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, present and future. *Trends in pharmacological sciences* **2004**, *25*, 193-200, doi:10.1016/j.tips.2004.02.007.
5. Buermans, H.P.; Vossen, R.H.; Anvar, S.Y.; Allard, W.G.; Guchelaar, H.J.; White, S.J.; den Dunnen, J.T.; Swen, J.J.; van der Straaten, T. Flexible and Scalable Full-Length CYP2D6 Long Amplicon PacBio Sequencing. *Human mutation* **2017**, 10.1002/humu.23166, doi:10.1002/humu.23166.
6. Qiao, W.; Yang, Y.; Sebra, R.; Mendiratta, G.; Gaedigk, A.; Desnick, R.J.; Scott, S.A. Long-Read Single Molecule Real-Time Full Gene Sequencing of Cytochrome P450-2D6. *Human mutation* **2016**, *37*, 315-323, doi:10.1002/humu.22936.
7. Owen, R.P.; Sangkuhl, K.; Klein, T.E.; Altman, R.B. Cytochrome P450 2D6. *Pharmacogenetics and genomics* **2009**, *19*, 559-562, doi:10.1097/FPC.0b013e32832e0e97.
8. Caudle, K.E.; Sangkuhl, K.; Whirl-Carrillo, M.; Swen, J.J.; Haidar, C.E.; Klein, T.E.; Gammal, R.S.; Relling, M.V.; Scott, S.A.; Hertz, D.L., et al. Standardizing CYP2D6 Genotype to Phenotype Translation: Consensus Recommendations from the Clinical Pharmacogenetics Implementation Consortium and Dutch Pharmacogenetics Working Group. *Clinical and translational science* **2020**, *13*, 116-124, doi:10.1111/cts.12692.
9. Pharmacogene Variation Consortium. CYP2D6 Allele Nomenclature. Availabe online: https://www.pharmvar.org/gene/CYP2D6
10. Gaedigk, A.; Ingelman-Sundberg, M.; Miller, N.A.; Leeder, J.S.; Whirl-Carrillo, M.; Klein, T.E. The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. *Clinical pharmacology and therapeutics* **2018**, *103*, 399-401, doi:10.1002/cpt.910.
11. Gaedigk, A.; Simon, S.D.; Pearce, R.E.; Bradford, L.D.; Kennedy, M.J.; Leeder, J.S. The CYP2D6 activity score: translating genotype information into a qualitative measure of phenotype. *Clinical pharmacology and therapeutics* **2008**, *83*, 234-242, doi:10.1038/sj.clpt.6100406.
12. Pharmacogenomics Knowledge Base. CYP2D6. Availabe online: https://www.pharmgkb.org/gene/PA128
13. 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. *Journal of personalized medicine* **2018**, *8*, doi:10.3390/jpm8020015.
14. Matthaei, J.; Brockmoller, J.; Tzvetkov, M.V.; Sehrt, D.; Sachse-Seeboth, C.; Hjelmborg, J.B.; Moller, S.; Halekoh, U.; Hofmann, U.; Schwab, M., et al. Heritability of metoprolol and torsemide pharmacokinetics. *Clinical pharmacology and therapeutics* **2015**, *98*, 611-621, doi:10.1002/cpt.258.
15. Micheli, J.E.; Chinn, L.W.; Shugarts, S.B.; Patel, A.; Martin, J.N.; Bangsberg, D.R.; Kroetz, D.L. Measuring the overall genetic component of nevirapine pharmacokinetics and the role of selected polymorphisms: towards addressing the missing heritability in pharmacogenetic phenotypes? *Pharmacogenetics and genomics* **2013**, *23*, 591-596, doi:10.1097/FPC.0b013e32836533a5.

16. Klein, K.; Zanger, U.M. Pharmacogenomics of Cytochrome P450 3A4: Recent Progress Toward the "Missing Heritability" Problem. *Frontiers in genetics* **2013**, *4*, 12, doi:10.3389/fgene.2013.00012.
17. Ingelman-Sundberg, M.; Mkrtchian, S.; Zhou, Y.; Lauschke, V.M. Integrating rare genetic variants into pharmacogenetic drug response predictions. *Human genomics* **2018**, *12*, 26, doi:10.1186/s40246-018-0157-3.
18. Hertz, D.L.; Snavely, A.C.; McLeod, H.L.; Walko, C.M.; Ibrahim, J.G.; Anderson, S.; Weck, K.E.; Magrinat, G.; Olajide, O.; Moore, S., et al. In vivo assessment of the metabolic activity of CYP2D6 diplotypes and alleles. *Br J Clin Pharmacol* **2015**, *80*, 1122-1130, doi:10.1111/bcp.12665.
19. Rajkomar, A.; Dean, J.; Kohane, I. Machine Learning in Medicine. *The New England journal of medicine* **2019**, *380*, 1347-1358, doi:10.1056/NEJMra1814259.
20. Zou, J.; Huss, M.; Abid, A.; Mohammadi, P.; Torkamani, A.; Telenti, A. A primer on deep learning in genomics. *Nature genetics* **2019**, *51*, 12-18, doi:10.1038/s41588-018-0295-5.
21. Lee, S.B.; Wheeler, M.M.; Patterson, K.; McGee, S.; Dalton, R.; Woodahl, E.L.; Gaedigk, A.; Thummel, K.E.; Nickerson, D.A. Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. *Genetics in medicine : official journal of the American College of Medical Genetics* **2019**, *21*, 361-372, doi:10.1038/s41436-018-0054-0.
22. McInnes, G.; Dalton, R.; Sangkuhl, K.; Whirl-Carrillo, M.; Lee, S.-b.; Altman, R.B.; Woodahl, E.L.J.B. Hubble2D6: A deep learning approach for predicting drug metabolic activity. **2019**, 684357.
23. Sanchez-Spitman, A.; Dezentje, V.; Swen, J.; Moes, D.; Bohringer, S.; Batman, E.; van Druten, E.; Smorenburg, C.; van Bochove, A.; Zeillemaker, A., et al. Tamoxifen Pharmacogenetics and Metabolism: Results From the Prospective CYPTAM Study. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* **2019**, 10.1200/jco.18.00307, Jco1800307, doi:10.1200/jco.18.00307.
24. Brauch, H.; Mürdter, T.E.; Eichelbaum, M.; Schwab, M. Pharmacogenomics of tamoxifen therapy. *Clinical chemistry* **2009**, *55*, 1770-1782, doi:10.1373/clinchem.2008.121756.
25. Sanchez-Spitman, A.B.; Swen, J.J.; Dezentje, V.O.; Moes, D.; Gelderblom, H.; Guchelaar, H.J. Clinical pharmacokinetics and pharmacogenetics of tamoxifen and endoxifen. *Expert Rev Clin Pharmacol* **2019**, *12*, 523-536, doi:10.1080/17512433.2019.1610390.
26. Schroth, W.; Winter, S.; Murdter, T.; Schaeffeler, E.; Eccles, D.; Eccles, B.; Chowbay, B.; Khor, C.C.; Tfayli, A.; Zgheib, N.K., et al. Improved Prediction of Endoxifen Metabolism by CYP2D6 Genotype in Breast Cancer Patients Treated with Tamoxifen. *Frontiers in pharmacology* **2017**, *8*, 582, doi:10.3389/fphar.2017.00582.
27. Sanchez Spitman, A.B.; Moes, D.; Gelderblom, H.; Dezentje, V.O.; Swen, J.J.; Guchelaar, H.J. Effect of CYP3A4*22, CYP3A5*3, and CYP3A combined genotypes on tamoxifen metabolism. *European journal of clinical pharmacology* **2017**, *73*, 1589-1598, doi:10.1007/s00228-017-2323-2.
28. Sanchez-Spitman, A.B.; Dezentjé, V.O.; Swen, J.J.; Moes, D.; Gelderblom, H.; Guchelaar, H.J. Genetic polymorphisms of 3'-untranslated region of SULT1A1 and their impact on tamoxifen metabolism and efficacy. *Breast cancer research and treatment* **2018**, *172*, 401-411, doi:10.1007/s10549-018-4923-7.
29. Helland, J.B.T.; Thomas, F.; Mellgren, G.; Swen, J.J.; Brauch, H, Gaedigk, A.; Hertz, D.L. Consensus CYP2D6 translation system improves explanation of tamoxifen bioactivation in analysis of 16 multi-racial cohorts [ABSTRACT]. In Proceedings of ASCPT 2020 Houston.
30. Kiyotani, K.; Mushiroda, T.; Imamura, C.K.; Tanigawara, Y.; Hosono, N.; Kubo, M.; Sasa, M.; Nakamura, Y.; Zembutsu, H. Dose-adjustment study of tamoxifen based on CYP2D6 genotypes in Japanese breast cancer patients. *Breast cancer research and treatment* **2012**, *131*, 137-145, doi:10.1007/s10549-011-1777-7.

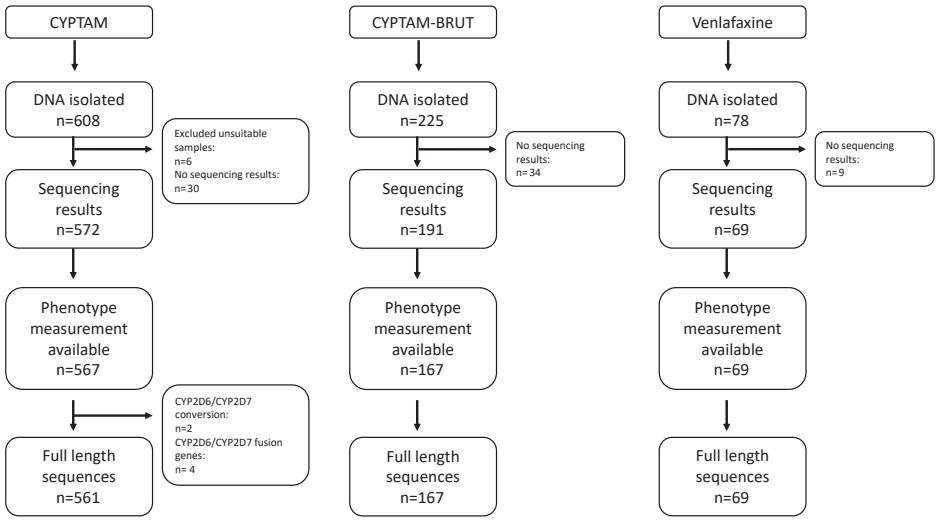
31. Neven, P.; Jongen, L.; Lintermans, A.; Van Asten, K.; Blomme, C.; Lambrechts, D.; Poppe, A.; Wildiers, H.; Dieudonne, A.S.; Brouckaert, O., et al. Tamoxifen Metabolism and Efficacy in Breast Cancer: A Prospective Multicenter Trial. *Clinical cancer research : an official journal of the American Association for Cancer Research* **2018**, *24*, 2312-2318, doi:10.1158/1078-0432.Ccr-17-3028.
32. Hicks, J.K.; Swen, J.J.; Gaedigk, A. Challenges in CYP2D6 phenotype assignment from genotype data: a critical assessment and call for standardization. *Current drug metabolism* **2014**, *15*, 218-232.
33. Muroi, Y.; Saito, T.; Takahashi, M.; Sakuyama, K.; Niinuma, Y.; Ito, M.; Tsukada, C.; Ohta, K.; Endo, Y.; Oda, A., et al. Functional characterization of wild-type and 49 CYP2D6 allelic variants for N-desmethyltamoxifen 4-hydroxylation activity. *Drug metabolism and pharmacokinetics* **2014**, *29*, 360-366.
34. Sangkuhl, K.; Stingl, J.C.; Turpeinen, M.; Altman, R.B.; Klein, T.E. PharmGKB summary: venlafaxine pathway. *Pharmacogenetics and genomics* **2014**, *24*, 62-72, doi:10.1097/fpc.0000000000000003.
35. Herman, D.; Dolzan, V.; Ingelman-Sundberg, M. Characterization of the novel defective CYP2C9*24 allele. *Drug metabolism and disposition: the biological fate of chemicals* **2007**, *35*, 831-834, doi:10.1124/dmd.106.013722.
36. Wang, J.; Sönnerborg, A.; Rane, A.; Josephson, F.; Lundgren, S.; Ståhle, L.; Ingelman-Sundberg, M. Identification of a novel specific CYP2B6 allele in Africans causing impaired metabolism of the HIV drug efavirenz. *Pharmacogenetics and genomics* **2006**, *16*, 191-198, doi:10.1097/01.fpc.0000189797.03845.90.
37. 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. *Pharmacogenetics and genomics* **2019**, *29*, 39-47, doi:10.1097/fpc.0000000000000363.
38. Wang, D.; Poi, M.J.; Sun, X.; Gaedigk, A.; Leeder, J.S.; Sadee, W. Common CYP2D6 polymorphisms affecting alternative splicing and transcription: long-range haplotypes with two regulatory variants modulate CYP2D6 activity. *Human molecular genetics* **2014**, *23*, 268-278, doi:10.1093/hmg/ddt417.
39. Sanchez-Spitman, A.B.; Moes, D.-J.A.; Gelderblom, H.; Dezentjé, V.O.; Swen, J.J.; Guchelaar, H.-J. The effect of rs5758550 on CYP2D6*2 phenotype and formation of endoxifen in breast cancer patients using tamoxifen. *Pharmacogenomics* **2017**, *18*, 1125-1132, doi:10.2217/pgs-2017-0080.
40. Sakuyama, K.; Sasaki, T.; Ujiie, S.; Obata, K.; Mizugaki, M.; Ishikawa, M.; Hiratsuka, M. Functional characterization of 17 CYP2D6 allelic variants (CYP2D6.2, 10, 14A-B, 18, 27, 36, 39, 47-51, 53-55, and 57). *Drug metabolism and disposition: the biological fate of chemicals* **2008**, *36*, 2460-2467, doi:10.1124/dmd.108.023242.
41. Saito, T.; Gutierrez Rico, E.M.; Kikuchi, A.; Kaneko, A.; Kumondai, M.; Akai, F.; Saigusa, D.; Oda, A.; Hirasawa, N.; Hiratsuka, M. Functional characterization of 50 CYP2D6 allelic variants by assessing primaquine 5-hydroxylation. *Drug metabolism and pharmacokinetics* **2018**, *33*, 250-257, doi:10.1016/j.dmpk.2018.08.004.
42. Department of medicine, Indiana University.The Flockhart Table. Availabe online: https://drug-interactions.medicine.iu.edu/Main-Table.aspx
43. Crews, K.R.; Gaedigk, A.; Dunnenberger, H.M.; Leeder, J.S.; Klein, T.E.; Caudle, K.E.; Haidar, C.E.; Shen, D.D.; Callaghan, J.T.; Sadhasivam, S., et al. Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450 2D6 genotype and codeine therapy: 2014 update. *Clinical pharmacology and therapeutics* **2014**, *95*, 376-382, doi:10.1038/clpt.2013.254.

44. Kringen, M.K.; Bråten, L.S.; Haslemo, T.; Molden, E. The Influence of Combined CYP2D6 and CYP2C19 Genotypes on Venlafaxine and O-Desmethylvenlafaxine Concentrations in a Large Patient Cohort. *J Clin Psychopharmacol* **2020**, *40*, 137-144, doi:10.1097/jcp.0000000000001174.
45. Dezentje, V.O.; Opdam, F.L.; Gelderblom, H.; Hartigh den, J.; Van der Straaten, T.; Vree, R.; Maartense, E.; Smorenburg, C.H.; Putter, H.; Dieudonne, A.S., et al. CYP2D6 genotype- and endoxifen-guided tamoxifen dose escalation increases endoxifen serum concentrations without increasing side effects. *Breast cancer research and treatment* **2015**, *153*, 583-590, doi:10.1007/s10549-015-3562-5.
46. Kisanga, E.R.; Gjerde, J.; Guerrieri-Gonzaga, A.; Pigatto, F.; Pesci-Feltri, A.; Robertson, C.; Serrano, D.; Pelosi, G.; Decensi, A.; Lien, E.A. Tamoxifen and metabolite concentrations in serum and breast cancer tissue during three dose regimens in a randomized preoperative trial. *Clinical cancer research : an official journal of the American Association for Cancer Research* **2004**, *10*, 2336-2343, doi:10.1158/1078-0432.ccr-03-0538.
47. Dong, A.N.; Ahemad, N.; Pan, Y.; Palanisamy, U.D.; Yiap, B.C.; Ong, C.E. Functional and structural characterisation of common cytochrome P450 2D6 allelic variants-roles of Pro34 and Thr107 in catalysis and inhibition. *Naunyn Schmiedebergs Arch Pharmacol* **2019**, *392*, 1015-1029, doi:10.1007/s00210-019-01651-0.
48. McInnes, G.; Dalton, R.; Sangkuhl, K.; Whirl-Carrillo, M.; Lee, S.-b.; Tsao, P.S.; Gaedigk, A.; Altman, R.B.; Woodahl, E.L. Transfer learning enables prediction of CYP2D6 haplotype function. *bioRxiv* **2020**, 10.1101/684357, 684357, doi:10.1101/684357.
49. Twesigomwe, D.; Wright, G.E.B.; Drögemöller, B.I.; da Rocha, J.; Lombard, Z.; Hazelhurst, S. A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: a focus on CYP2D6 genotyping. *NPJ genomic medicine* **2020**, *5*, 30, doi:10.1038/s41525-020-0135-2.
50. Pan, X.; Ning, M.; Jeong, H. Transcriptional Regulation of CYP2D6 Expression. *Drug metabolism and disposition: the biological fate of chemicals* **2017**, *45*, 42-48, doi:10.1124/dmd.116.072249.
51. He, Z.X.; Chen, X.W.; Zhou, Z.W.; Zhou, S.F. Impact of physiological, pathological and environmental factors on the expression and activity of human cytochrome P450 2D6 and implications in precision medicine. *Drug metabolism reviews* **2015**, *47*, 470-519, doi:10.3109/03602532.2015.1101131.
52. Ning, M.; Duarte, J.D.; Stevison, F.; Isoherranen, N.; Rubin, L.H.; Jeong, H. Determinants of Cytochrome P450 2D6 mRNA Levels in Healthy Human Liver Tissue. *Clinical and translational science* **2019**, *12*, 416-423, doi:10.1111/cts.12632.
53. Genotype Tissue Expression project. GTEx Portal. Availabe online: https://gtexportal.org/home/
54. Qin, S.; Eugene, A.R.; Liu, D.; Zhang, L.; Neavin, D.; Biernacka, J.M.; Yu, J.; Weinshilboum, R.M.; Wang, L. Dual Roles for the TSPYL Family in Mediating Serotonin Transport and the Metabolism of Selective Serotonin Reuptake Inhibitors in Patients with Major Depressive Disorder. *Clinical pharmacology and therapeutics* **2020**, *107*, 662-670, doi:10.1002/cpt.1692.
55. Sanchez-Spitman, A.B.; Swen, J.J.; Dezentjé, V.O.; Moes, D.; Gelderblom, H.; Guchelaar, H.J. Effect of CYP2C19 genotypes on tamoxifen metabolism and early-breast cancer relapse. *Scientific reports* **2021**, *11*, 415, doi:10.1038/s41598-020-79972-x.
56. Martinez-Matilla, M.; Blanco-Verea, A.; Santori, M.; Ansede-Bermejo, J.; Ramos-Luis, E.; Gil, R.; Bermejo, A.M.; Lotufo-Neto, F.; Hirata, M.H.; Brisighelli, F., et al. Genetic susceptibility in pharmacodynamic and pharmacokinetic pathways underlying drug-induced arrhythmia and sudden unexplained deaths. *Forensic Sci Int Genet* **2019**, *42*, 203-212, doi:10.1016/j.fsigen.2019.07.010.

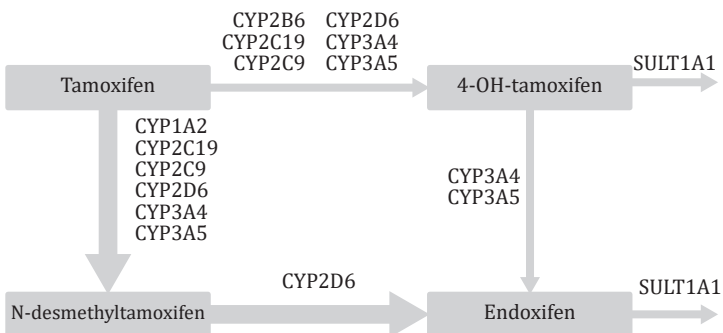
57. Slanař, O.; Hronová, K.; Bartošová, O.; Šíma, M. Recent advances in the personalized treatment of estrogen receptor-positive breast cancer with tamoxifen: a focus on pharmacogenomics. *Expert opinion on drug metabolism & toxicology* **2020**, 10.1080/17425255.2021.1865310, 1-15, doi:10.1080/17425255.2021.1865310.
58. Shah, R.R.; Smith, R.L. Addressing phenoconversion: the Achilles' heel of personalized medicine. *British journal of clinical pharmacology* **2015**, *79*, 222-240, doi:10.1111/bcp.12441.
59. Shah, R.R.; Smith, R.L. Inflammation-induced phenoconversion of polymorphic drug metabolizing enzymes: hypothesis with implications for personalized medicine. *Drug metabolism and disposition: the biological fate of chemicals* **2015**, *43*, 400-410, doi:10.1124/dmd.114.061093.
60. Zhou, Y.; Ingelman-Sundberg, M.; Lauschke, V.M. Worldwide distribution of cytochrome P450 alleles: A meta-analysis of population-scale sequencing projects. *Clinical pharmacology and therapeutics* **2017**, *102*, 688-700, doi:10.1002/cpt.690.
61. Kozyra, M.; Ingelman-Sundberg, M.; Lauschke, V.M. Rare genetic variants in cellular transporters, metabolic enzymes, and nuclear receptors can be important determinants of interindividual differences in drug response. *Genetics in medicine : official journal of the American College of Medical Genetics* **2017**, *19*, 20-29, doi:10.1038/gim.2016.33.
62. Pacific Biosciences. Availabe online: https://www.pacb.com
63. Gaedigk, A.; Ndjountche, L.; Divakaran, K.; Dianne Bradford, L.; Zineh, I.; Oberlander, T.F.; Brousseau, D.C.; McCarver, D.G.; Johnson, J.A.; Alander, S.W., et al. Cytochrome P4502D6 (CYP2D6) gene locus heterogeneity: characterization of gene duplication events. *Clinical pharmacology and therapeutics* **2007**, *81*, 242-251, doi:10.1038/sj.clpt.6100033.
64. Gaedigk, A.; Coetsee, C. The CYP2D6 gene locus in South African Coloureds: unique allele distributions, novel alleles and gene arrangements. *European journal of clinical pharmacology* **2008**, *64*, 465-475, doi:10.1007/s00228-007-0445-7.
65. Integrated DNA technologies. Availabe online: https://eu.idtdna.com/pages
66. Takara Bio Availabe online: https://www.takarabio.com
67. Roche sequencing. KAPA long range PCR kits. Availabe online: https://sequencing.roche.com/en/products-solutions/by-category/pcr/kapa-long-range-pcr-kits/ordering.html
68. Pacific Bioscience. Sequel Systems. Availabe online: https://www.pacb.com/products-and-services/sequel-system/
69. McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.; Thormann, A.; Flicek, P.; Cunningham, F. The Ensembl Variant Effect Predictor. *Genome biology* **2016**, *17*, 122, doi:10.1186/s13059-016-0974-4.
70. Clinical Pharmacogenetics Implementation Consortium. CPIC-guidelines. Availabe online: https://cpicpgx.org/
71. Royal Dutch Pharmacy Association. Pharmacogenetics. Availabe online: www.knmp.nl/farmacogenetica
72. Khalaj, Z.; Baratieh, Z.; Nikpour, P.; Schwab, M.; Schaeffeler, E.; Mokarian, F.; Khanahmad, H.; Salehi, R.; Murdter, T.E.; Salehi, M. Clinical Trial: CYP2D6 Related Dose Escalation of Tamoxifen in Breast Cancer Patients With Iranian Ethnic Background Resulted in Increased Concentrations of Tamoxifen and Its Metabolites. *Frontiers in pharmacology* **2019**, *10*, 530, doi:10.3389/fphar.2019.00530.
73. Kashuba, A.D.; Nafziger, A.N.; Kearns, G.L.; Leeder, J.S.; Shirey, C.S.; Gotschall, R.; Gaedigk, A.; Bertino, J.S., Jr. Quantification of intraindividual variability and the influence of menstrual cycle phase on CYP2D6 activity as measured by dextromethorphan phenotyping. *Pharmacogenetics* **1998**, *8*, 403-410, doi:10.1097/00008571-199810000-00005.

74. Alfaro, C.L.; Lam, Y.W.; Simpson, J.; Ereshefsky, L. CYP2D6 inhibition by fluoxetine, paroxetine, sertraline, and venlafaxine in a crossover study: intraindividual variability and plasma concentration correlations. *Journal of clinical pharmacology* **2000**, *40*, 58-66, doi:10.1177/00912700022008702.
75. Fotoohi, A.K.; Karim, H.; Lafolie, P.; Pohanka, A.; Östervall, J.; Hatschek, T.; Vitols, S. Pronounced Interindividual But Not Intraindividual Variation in Tamoxifen and Metabolite Levels in Plasma During Adjuvant Treatment of Women With Early Breast Cancer. *Ther Drug Monit* **2016**, *38*, 239-245, doi:10.1097/FTD.0000000000000257.
76. Oscarson, M.; Hidestrand, M.; Johansson, I.; Ingelman-Sundberg, M. A combination of mutations in the CYP2D6*17 (CYP2D6Z) allele causes alterations in enzyme function. *Mol Pharmacol* **1997**, *52*, 1034-1040, doi:10.1124/mol.52.6.1034.
77. Kronbach, T.; Mathys, D.; Gut, J.; Catin, T.; Meyer, U.A. High-performance liquid chromatographic assays for bufuralol 1'-hydroxylase, debrisoquine 4-hydroxylase, and dextromethorphan O-demethylase in microsomes and purified cytochrome P-450 isozymes of human liver. *Anal Biochem* **1987**, *162*, 24-32, doi:10.1016/0003-2697(87)90006-6.

Supplementary material

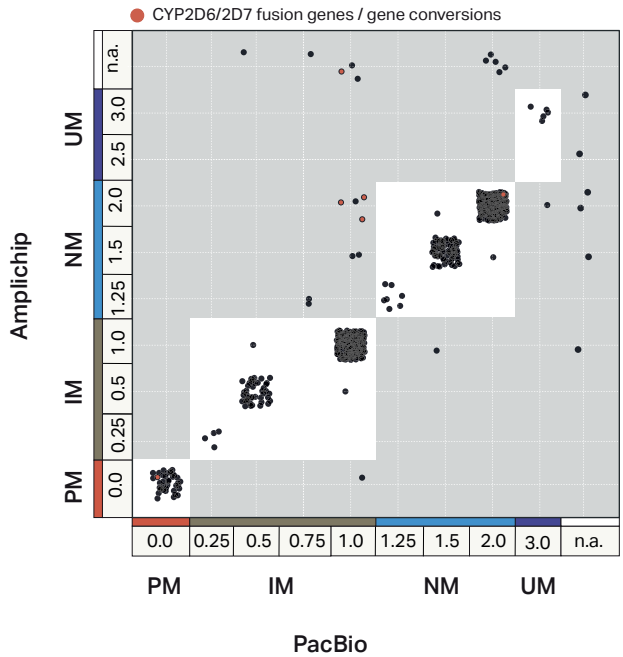


Supplementary Figure S5.1: Flowchart of study cohorts
Samples were selected based on availability of remaining DNA. Samples were excluded if patients no longer wanted to be part of the main study or were double included (total: n=6 in CYPTAM). All samples were sequenced for *CYP2D6* with PacBio SMRT sequencing. For neural network training and predictions only samples with full length CYP2D6 sequences available and with clinical phenotype data were included.

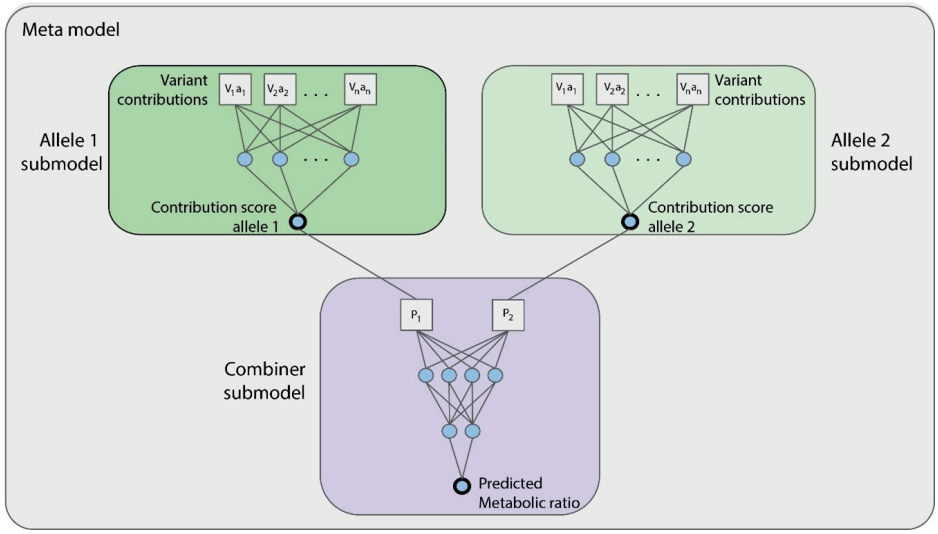


Proxy for **CYP2D6 Enzyme activity**
LN(Endoxifen/Desmethyltamoxifen)

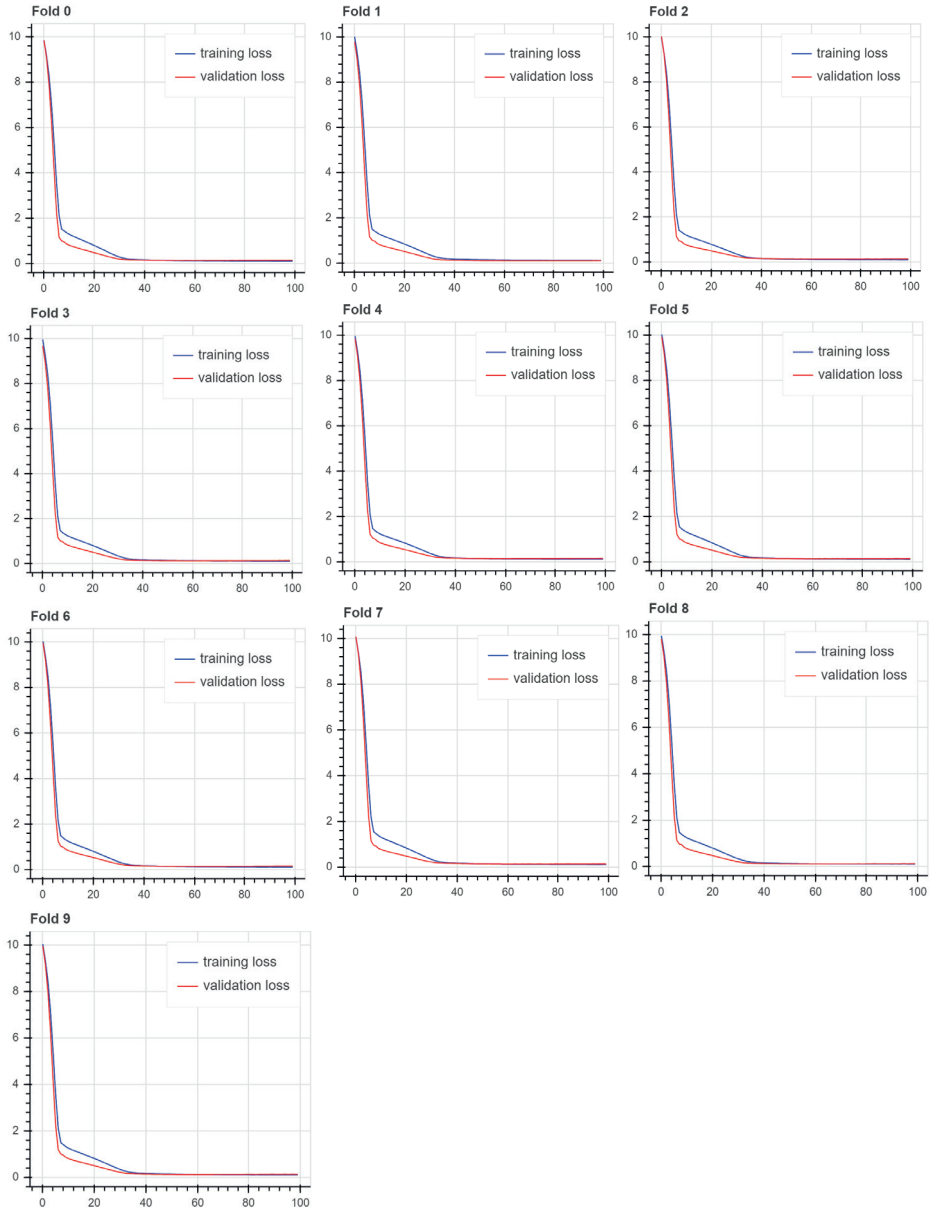
Supplementary Figure S5.2: Metabolic pathway of tamoxifen
Tamoxifen is first metabolized into desmethyltamoxifen and 4-hydroxytamoxifen, followed by a conversion of these metabolites into endoxifen. The path through desmethyltamoxifen is the predominant pathway, responsible for the majority of the endoxifen formation. CYP2D6 plays a key role in all metabolic conversions to endoxifen. Depicted is the core metabolism of tamoxifen into its most active metabolite endoxifen.



Supplementary Figure S5.3: Concordance Amplichip- and PacBio-based phenotype predictions
For all CYPTAM individuals phenotype predictions were made, based on Amplichip genotyping and PacBio SMRT-sequencing genotype calling, according to Gene Activity Scores (GAS) and the Dutch Pharmacogenetics Working Group and Clinical implementation consortiums consensus guideline. Concordance: Kappa-coefficient 0.94, p<0.0001.



Supplementary Figure S5.4: Neural network design
The neural network model consists of two parts. The first part consists of two allele submodels which train as one. These models produce a contribution score per allele. The second part, the combiner submodel, combines the contribution scores into a predicted metabolic ratio. Variant contributions reflect the impact of variants on enzyme function, the contribution scores are normalized to represent gene activity scores, the predicted metabolic ratio serves as a proxy for CYP2D6 enzyme activity.



Supplementary Figure S5.5: 10-fold crossvalidtion with internal hold-out
25% of the data was used for the validation set. No devation between training and validation loss was observed up to 100 epochs. No signs of overfitting were observed.

Supplementary Table S5.1: *-allele haplotype frequencies for all cohorts
All haplotypes are based on all variants observed in the CYP2D6 locus, including CYP2D6/2D7 conversions and fusion genes. CYPTAM-cohort is the training cohort, CYPTAM-BRUT the first validation cohort with tamoxifen as the CYP2D6 substrate used, Venlafaxine is the second replication cohort with individuals using the CYP2D6 substrate venlafaxine. Haplotype translations are based on the PharmGKB variant to haplotype translations.

Haplotype	CYPTAM		CYPTAM-BRUT		Venlafaxine	
	n	%	n	%	n	%
*1	357	32.8	78	30.7	33	23.91
*108	3	0.27	1	0.39		
*10A			1	0.39		
*10B			1	0.39		
*10D	18	1.6	3	1.2	2	1.45
*15					1	0.72
*17	2	0.18				
*1B	15	1.3	2	0.79		
*1D	2	0.18			1	0.72
*1E	6	0.53				
*1xN	6	0.53	1	0.39		
*2	4	0.36				
*22	5	0.45				
*27			1	0.39		
*2A	181	16.1	51	20.1	24	17.39
*2AxN	1	0.09	1	0.39	2	1.45
*2D			2	0.79	1	0.72
*2M	1	0.09				
*2xN	1	0.09				
*31	1	0.09				
*33	13	1.2	1	0.39	2	1.45
*34			1	0.39		
*35A	59	5.3	18	7.1	8	5.80
*35AxN	2	0.18				
*39					1	0.72
*3A	21	1.9	3	1.2	7	5.07
*41	92	8.2	24	9.4	16	11.59
*41xN	1	0.09				
*4A	221	19.7	33	13.0	21	15.22
*4AxN	2	0.18	2	0.79	1	0.72
*4B					1	0.72
*4D	5	0.45	8	3.1	4	2.9
*4H	1	0.09				
*4J			1	0.39		
*5	45	4.0	9	3.5	8	5.80
*59	6	0.54				
*6A						
*6B	15	1.3	4	1.6	2	1.45
*7			1	0.39		
*9	36	3.2	7	2.8	3	2.17

Supplementary Table S5.2: Bufuralol incubation results
Site mutagenesis on HEK cells was performed on pCMV4 CYP2D6*1 plasmid with QuikChange II Site-Directed Mutagenesis Kit (Agilent). The D337N exchange was performed using pCMV4 CYP2D6*2 as template. HEK cells were incubated with bufuralol for 5 hours, upon which the metabolism rate was assessed by measuring bufuralol and metabolites. Results are normalised to the average activity of *1, which is set at 1.0.

Genotype	September 2019		December 2019		January 2020		Average
*1	0.927845	1.072155	0.952799	1.047201	1.027143	0.972857	1
*2	0.780427	0.806543	0.520376	0.474208	0.455362		0.607383
*2+D337N	0.438914	0.66233	0	0	0.018975		0.224044
R330P	0	0	0	0	0.007519	0.047079	0,0091
G42E	0	0	0	0	0.009874	0.001144	0.001836
F120I	3.439576	4.910881	5.878972	5.374999	3.100804	2.457602	4.193805

Supplementary Table S5.3: Primer sequences for three separate PCR reactions
Full-length PCR: yielding one full length CYP2D6 sequence. Duplex PCR (*5): yielding an internal control fragment for all samples and a deletion fragment if a CYP2D6 deletion is present. Triplex PCR: Yielding a full-length fragment as well as a duplication fragment in the presence of a CYP2D6 duplication and/or a hybrid fragment in the presence of a CYP2D6/CYP2D7 fusion gene.

Name	Primer sequence (5'–3')
Full-length PCR	
Fragment A Forward	ATGGCAGCTGCCATACAATCCACCTG
Fragment A Reverse	CGACTGAGCCCTGGGAGGTAGGTAG
Duplex PCR (*5)	
Fragment *5-forward	CTCCAGCCTCCACCAGTCCAG
Fragment *5-reverse	CAGGCATGAGCTAAGGCACCCAGAC
IC-forward	GCATGCACAGCTCAGCACTGC
IC-reverse	GCCACCCTGATGTCTCAGTTTCG
Triplex PCR	
Fragment A Forward	ATGGCAGCTGCCATACAATCCACCTG
Fragment A Reverse	CGACTGAGCCCTGGGAGGTAGGTAG
Fragment B - Forward	CCATGGAAGCCCAGGACTGAGC
Fragment B - Reverse	CGGCAGTGGTCAGCTAATGAC
Fragment H - Forward	TCCGACCAGGCCTTTCTACCAC

IC: Internal control.

Supplementary Table S5.4: HEK cell mutagenesis primers

Site mutagenesis on HEK cells was performed on pCMV4 CYP2D6*1 plasmid with QuikChange II Site-Directed Mutagenesis Kit (Agilent). The D337N exchange was performed using pCMV4 CYP2D6*2 as template.

Nucl. mutation	aa mutation	Primer FWD 5´- 3	Primer REV 5´- 3
125 G>A	G42E	CCTGCCACTGCCCG**A**GCTGGGCAACCTGCT	AGCAGGTTGCCCAGC**T**CGGGCAGTGGCAGG
1009 G>A	D337N	CAACAGGAGATCGAC**A**ACGTGATAGGGCAGG	CCTGCCCTATCACGT**T**GTCGATCTCCTGTTG
1611 T>A	F120E	CGTTCCCAAGGGGTG**A**TCCTGGCGCGCTATG	CATAGCGCGCCAGGA**T**CACCCCTTGGGAACG
3160 G>C	R330P	CGGATGTGCAGCGCC**C**TGTCCAACAGGAGAT	ATCTCCTGTTGGACA**G**GGCGCTGCACATCCG

Supplementary Data File S5.1: Variant frequencies and predicted contributions

All variants (mapped to GRCh38) which are identified in all individuals included in this study. Variant Effect Prediction (VEP) based on the most severe effect are included. Variants were included in the neural network if they were part of *-allele nomenclature or if they were classified as missense, frameshift or splice region variants. Based on the neural network each variant was assigned a variant contribution score scaled to -1.0 for a deletion and 0 for no effect compared to wildtype. The number of associated eQTLs is extracted from GTEx, no variants were part of eQTLs in the liver.

Variant (NC_000022.11:g.)	CYPTAM	CYPTAM-BRUT	Venlafaxine-cohort	Total	Minor allele frequency (1,596 alleles)	Included neural network	Predicted contribution to allele activity based on neural network
42126069A>C	600	187	81	868	54.4%	no	
42126074G>A	2	0	0	2	0.1%	no	
42126079C>T	2	0	0	2	0.1%	no	
42126136_42126138delTGT	348	126	51	525	32.9%	no	
42126136delT	3	0	0	3	0.2%	no	
42126138_42126139delTG	6	0	0	6	0.4%	no	
42126201_42126202insG	0	1	0	1	0.1%	no	
42126310C>T	348	127	52	527	33.0%	yes	-0.043653817811100566
42126343_42126344insG	1	0	0	1	0.1%	no	
42126347T>C	1	0	0	1	0.1%	no	
42126389_42126390insA	0	2	0	2	0.1%	no	
42126390G>A	247	60	29	336	21.1%	yes	-0.027091746508636685
42126396_42126397delCT	2	0	0	2	0.1%	no	
42126417G>A	1	0	0	1	0.1%	no	
42126462G>A	14	2	0	16	1.0%	no	
42126552G>T	0	0	2	2	0.1%	no	
42126611C>G	601	184	81	866	54.3%	yes	0.02252113634360636
42126611delC	0	1	0	1	0.1%	no	
42126676G>A	5	2	0	7	0.4%	yes	-0.09213514232176667
42126749C>T	1	0	0	1	0.1%	yes	-0.109389898471713
42126763G>T	2	0	0	2	0.1%	no	

Supplementary Data File S5.1 continues on next page.

Supplementary Data File S5.1: ***Continued***

Variant (NC_000022.11:g.)	CYPTAM	CYPTAM-BRUT	Venlafaxine-cohort	Total	Minor allele frequency (1,596 alleles)	Included neural network	Predicted contribution to allele activity based on neural network
42126798A>G	0	1	0	1	0.1%	no	
42126914C>G	1	0	0	1	0.1%	yes	0.133014501516827
42126938C>T	1	1	0	2	0.1%	yes	-0.0489283204986387
42126940delC	1	0	0	1	0.1%	yes	-0.218086598672027
42126944_42126945insT	0	0	1	1	0.1%	no	
42126944C>T	0	0	1	1	0.1%	no	
42126963C>T	16	2	0	18	1.1%	yes	-0.00251656770706176
42127001G>A	353	126	51	530	33.2%	yes	-0.058850394544319715
42127019G>T	1	0	0	1	0.1%	no	
42127173C>T	1	0	0	1	0.1%	no	
42127207C>T	353	126	51	530	33.2%	yes	-0.026299823547962564
42127209T>C	250	61	30	341	21.4%	yes	-0.046035356570000484
42127356G>T	29	5	3	37	2.3%	yes	-0.036148767544496224
42127385C>A	1	0	0	1	0.1%	no	
42127398A>G	5	9	4	18	1.1%	no	
42127407T>G	603	184	82	869	54.4%	yes	-0.08095961630167535
42127458G>A	1	0	0	1	0.1%	yes	0.174335300922393
42127526_42127527insT	0	1	0	1	0.1%	no	
42127526C>T	2	1	0	3	0.2%	yes	0.130049116392221
42127556T>C	3	1	0	4	0.3%	yes	-0.08524450659751888
42127565T>C	3	1	0	4	0.3%	yes	-0.165771216154098
42127611C>T	3	1	1	5	0.3%	yes	-0.16890266655038766
42127631C>G	1	2	0	3	0.2%	yes	-0.496762812137603
42127634C>A	1	0	0	1	0.1%	yes	NA
42127644C>T	2	0	0	2	0.1%	no	

42127677_42127687delGTCCGGCCCTG	1	0	0	1	0.1%	no	
42127694_42127695delCT	1	0	0	1	0.1%	no	
42127698T>C	1	0	0	1	0.1%	no	
42127700G>T	1	0	0	1	0.1%	no	
42127707A>G	1	0	0	1	0.1%	no	
42127718G>A	1	0	0	1	0.1%	no	
42127721C>T	1	0	0	1	0.1%	no	
42127734T>G	1	0	0	1	0.1%	no	
42127740T>C	1	0	0	1	0.1%	no	
42127743A>C	1	0	0	1	0.1%	no	
42127753T>G	1	0	0	1	0.1%	no	
42127755G>A	1	0	0	1	0.1%	no	
42127761C>T	1	0	0	1	0.1%	yes	0.0785754323005676
42127774A>T	1	0	0	1	0.1%	no	
42127778T>C	1	0	0	1	0.1%	no	
42127779T>C	1	0	0	1	0.1%	no	
42127782T>G	1	0	0	1	0.1%	no	
42127791G>A	1	0	0	1	0.1%	no	
42127792C>G	1	0	0	1	0.1%	no	
42127803C>T	93	34	16	143	9.0%	yes	-0.37951637920601106
42127811G>A	4	0	2	6	0.4%	no	
42127813C>T	1	0	0	1	0.1%	no	
42127820A>C	1	0	0	1	0.1%	no	
42127821C>T	1	0	0	1	0.1%	no	
42127824T>G	1	0	0	1	0.1%	no	
42127825T>G	1	0	0	1	0.1%	no	
42127826T>C	1	0	0	1	0.1%	no	
42127832A>C	1	0	0	1	0.1%	no	

Supplementary Data File S5.1 continues on next page.

Supplementary Data File S5.1: ***Continued***

Variant (NC_000022.11:g.)	CYPTAM	CYPTAM-BRUT	Venlafaxine-cohort	Total	Minor allele frequency (1,596 alleles)	Included neural network	Predicted contribution to allele activity based on neural network
42127852C>T	6	0	0	6	0.4%	yes	-0.46053318055295134
42127853G>A	1	0	0	1	0.1%	yes	NA
42127855A>G	1	0	0	1	0.1%	no	
42127856T>G	0	1	0	1	0.1%	no	
42127940_42127941insA	0	1	0	1	0.1%	no	
42127941G>A	353	127	51	531	33.3%	yes	-0.05377423686868765
42127973T>C	0	1	0	1	0.1%	no	
42127996G>A	1	0	0	1	0.1%	no	
42128042A>G	1	0	0	1	0.1%	no	
42128071C>G	5	4	1	10	0.6%	no	
42128088G>A	1	0	0	1	0.1%	no	
42128130C>T	4	0	0	4	0.3%	yes	0.022598231385327648
42128136C>T	0	1	0	1	0.1%	no	
42128176_42128178delTCT	37	11	3	51	3.2%	yes	-0.37228316068649164
42128176delT	1	0	0	1	0.1%	yes	NA
42128216G>T	2	0	1	3	0.2%	yes	-0.0146875381469726
42128218delG	27	0	0	27	1.7%	yes	0.047781497582548245
42128242delT	22	4	7	33	2.1%	yes	-1
42128251_42128254delTTAG	1	0	1	2	0.1%	yes	-0.54435521364212
42128308C>A	13	1	2	16	1.0%	yes	0.019396603107452302
42128321A>G	4	3	2	9	0.6%	yes	0.228823016128404
42128327G>T	0	1	0	1	0.1%	no	
42128438C>T	2	1	1	4	0.3%	no	
42128500C>T	13	0	0	13	0.8%	yes	-0.1153512736167856
42128519G>A	1	0	0	1	0.1%	no	

42128631C>T	1	0	0	1	0.1%	no	
42128693_42128694insC	0	1	0	1	0.1%	no	
42128694T>C	250	59	30	339	21.2%	yes	0.05622132786746358
42128736C>T	1	0	0	1	0.1%	no	
42128741G>A	1	0	1	2	0.1%	no	
42128793A>G	6	1	0	7	0.4%	no	
42128815C>T	15	6	2	23	1.4%	yes	-0.327434092760086
42128858C>G	0	1	0	1	0.1%	no	
42128922A>G	6	0	0	6	0.4%	yes	0.10502272844314499
42128945C>T	230	55	27	312	19.5%	yes	-0.22914504557519727
42128946_42128947insT	0	1	0	1	0.1%	no	
42129084delA	15	7	2	24	1.5%	yes	-0.616389781236648
42129087G>C	2	0	1	3	0.2%	yes	0.2265237666148455
42129130C>G	603	186	79	868	54.4%	yes	-0.15171797035812631
42129174C>A	1	0	0	1	0.1%	yes	0.237182915210723
42129178G>A	4	1	0	5	0.3%	no	
42129180A>T	2	0	0	2	0.1%	yes	0.5293659449718591
42129184C>T	0	1	0	1	0.1%	no	
42129225A>C	0	0	1	1	0.1%	no	
42129228A>G	0	0	1	1	0.1%	no	
42129303T>G	2	0	0	2	0.1%	no	
42129319_42129320delTT	1	0	0	1	0.1%	no	
42129351A>G	1	0	0	1	0.1%	no	
42129388G>A	6	1	0	7	0.4%	no	
42129414G>C	4	1	0	5	0.3%	no	
42129436A>G	2	0	0	2	0.1%	no	
42129450delG	1	0	0	1	0.1%	no	
42129461G>A	1	0	0	1	0.1%	no	

Supplementary Data File S5.1 continues on next page.

Supplementary Data File S5.1: ***Continued***

Variant (NC_000022.11:g.)	CYPTAM	CYPTAM-BRUT	Venlafaxine-cohort	Total	Minor allele frequency (1,596 alleles)	Included neural network	Predicted contribution to allele activity based on neural network
42129498C>T	1	0	0	1	0.1%	no	
42129545G>A	16	6	2	24	1.5%	no	
42129581T>G	1	0	0	1	0.1%	no	
42129587C>T	1	0	0	1	0.1%	no	
42129614C>G	2	0	0	2	0.1%	no	
42129623C>T	80	19	10	109	6.8%	no	
42129643G>C	2	0	0	2	0.1%	no	
42129722C>T	1	0	0	1	0.1%	no	
42129726A>C	1	0	0	1	0.1%	no	
42129731T>C	1	0	0	1	0.1%	yes	-0.243858930342826
42129754G>A	23	13	6	42	2.6%	yes	0.058333239961291326
42129770G>A	2	0	0	2	0.1%	yes	0.14140502977480102
42129793G>A	0	0	1	1	0.1%	no	
42129795_42129796insC	0	1	0	1	0.1%	no	
42129796G>C	225	46	23	294	18.4%	yes	0.1486052139837259
42129809T>C	225	48	24	297	18.6%	yes	-0.25058796488421997
42129819G>T	225	48	23	296	18.5%	yes	-0.036173526027646376
42129950A>C	610	188	80	878	55.0%	yes	-0.29604015773312636
42129950delA	1	0	0	1	0.1%	no	
42130047G>C	578	174	74	826	51.8%	yes	0.04334876873857956
42130180C>T	1	0	0	1	0.1%	no	
42130343T>G	0	1	0	1	0.1%	no	
42130469C>T	1	0	1	2	0.1%	no	
42130475T>C	1	0	1	2	0.1%	no	
42130482C>A	603	188	79	870	54.5%	yes	0.10251049786610479

42130522G>A	44	15	7	66	4.1%	no	
42130538C>A	0	0	1	1	0.1%	no	
42130547T>C	354	127	52	533	33.4%	yes	0.020412143103732624
42130552G>A	1	0	0	1	0.1%	no	
42130559T>G	354	127	52	533	33.4%	yes	-0.016224358148486787
42130560C>G	354	127	52	533	33.4%	yes	-0.02308920692216708
42130565A>G	354	127	52	533	33.4%	yes	0.06461922083170608
42130569G>C	354	127	52	533	33.4%	yes	-0.24267589335388073
42130571G>T	354	127	52	533	33.4%	yes	0.1175802367055315
42130578C>G	354	127	52	533	33.4%	yes	-0.005272804071269889
42130594_42130595insT	0	2	0	2	0.1%	no	
42130655_42130656insA	0	0	2	2	0.1%	no	
42130657G>A	1	0	0	1	0.1%	no	
42130667C>T	1	0	0	1	0.1%	yes	-0.521055400371551
42130692G>A	248	60	29	337	21.1%	yes	-0.07567874934073372
42130710G>A	5	1	0	6	0.4%	yes	-0.103839337825775
42130761_42130762insT	0	0	2	2	0.1%	no	
42130761C>T	61	28	8	97	6.1%	yes	-0.027715351754274285
42130773C>T	2	0	1	3	0.2%	yes	-0.09531869985836064
42130929T>C	1	0	0	1	0.1%	yes	0.10511073372503
42131000G>A	0	1	0	1	0.1%	no	
42131062T>A	1	0	0	1	0.1%	no	
42131114A>G	5	1	0	6	0.4%	no	
42131156C>T	18	8	0	26	1.6%	no	
42131222G>A	1	1	0	2	0.1%	yes	-0.0395302993694236
42131379C>T	11	0	1	12	0.8%	no	
42131469C>T	350	127	51	528	33.1%	yes	-0.02665604532520481
42131493C>T	0	0	1	1	0.1%	no	

Supplementary Data File S5.1 continues on next page.

Supplementary Data File S5.1: ***Continued***

Variant (NC_000022.11:g.)	CYPTAM	CYPTAM-BRUT	Venlafaxine-cohort	Total	Minor allele frequency (1,596 alleles)	Included neural network	Predicted contribution to allele activity based on neural network
42131531G>A	354	127	51	532	33.3%	yes	-0.013184694660057422
42131546_42131547delTC	1	0	0	1	0.1%	yes	-0.0190738309697214
42131610G>C	2	0	0	2	0.1%	no	
42131631A>T	3	1	0	4	0.3%	no	
42131681C>T	0	1	0	1	0.1%	no	
42131775C>T	0	1	1	2	0.1%	no	
42131791C>T	248	60	29	337	21.1%	yes	-0.22362225330200997
42131894delT	898	2	4	904	56.6%	no	
42131908C>T	1	0	0	1	0.1%	no	
42131914_42131915delCT	0	0	1	1	0.1%	no	
42131915T>G	1	0	0	1	0.1%	no	
42131927C>A	3	1	0	4	0.3%	no	
42131990delC	15	1	1	17	1.1%	no	
42132026T>C	604	187	80	871	54.6%	yes	0.30988314411907114
42132027_42132028insC	0	1	0	1	0.1%	no	
42132045_42132049delTTTTT	1	0	0	1	0.1%	no	
42132046_42132049delTTTT	3	0	0	3	0.2%	no	
42132047_42132049delTTT	6	0	0	6	0.4%	no	
42132048_42132049delTT	244	40	6	290	18.2%	no	
42132049_42132050insT	109	34	19	162	10.2%	no	
42132049_42132050insTT	194	55	17	266	16.7%	no	
42132049_42132050insTTT	86	57	28	171	10.7%	no	
42132049_42132050insTTTT	0	13	5	18	1.1%	no	
42132049delT	259	79	33	371	23.2%	no	
42132138C>G	0	0	1	1	0.1%	no	

42132217G>A	250	62	30	342	21.4%	yes	-0.189309422790393
42132375G>C	253	93	34	380	23.8%	yes	0.06632569069120157
42132377delG	1	0	0	1	0.1%	no	
42132411C>T	1	0	0	1	0.1%	no	
42132561_42132562insT	0	1	0	1	0.1%	no	
42132561C>T	349	126	50	525	32.9%	yes	0.06237363234918848
42132577G>C	0	1	0	1	0.1%	no	
42132589C>T	0	1	0	1	0.1%	no	
42132590G>A	0	1	0	1	0.1%	no	
Del	39	10	8	57	3.6%	yes	-1
Dup	13	4	3	20	1.3%	yes	0.2267

Supplementary Data File S5.1 continues on next page.

Supplementary Data File S5.1: ***Continued***

Variant (NC_000022.11:g.)	Consequence	rs-number	Sift prediction	Sift score	Polyphen prediction	Polyphen score	Number of associated eQTLs according to GTEx
42126069A>C	downstream_gene_variant	rs35028622	-	-	-	-	46
42126074G>A	downstream_gene_variant	rs77827855	-	-	-	-	-
42126079C>T	downstream_gene_variant	rs4078249	-	-	-	-	-
42126136_42126138delTGT	downstream_gene_variant	rs71184866	-	-	-	-	-
42126136delT	downstream_gene_variant		-	-	-	-	-
42126138_42126139delTG	downstream_gene_variant		-	-	-	-	-
42126201_42126202insG	downstream_gene_variant		-	-	-	-	-
42126310C>T	downstream_gene_variant	rs12169962	-	-	-	-	24
42126343_42126344insG	downstream_gene_variant	rs1200940710	-	-	-	-	-
42126347T>C	downstream_gene_variant	rs148648640	-	-	-	-	-
42126389_42126390insA	downstream_gene_variant		-	-	-	-	-
42126390G>A	downstream_gene_variant	rs28371738	-	-	-	-	36
42126396_42126397delCT	downstream_gene_variant	rs202032066	-	-	-	-	-
42126417G>A	downstream_gene_variant	rs550569325	-	-	-	-	-
42126462G>A	downstream_gene_variant	rs28371737	-	-	-	-	-
42126552G>T	3_prime_UTR_variant	rs746767060	-	-	-	-	-
42126611C>G	missense_variant	rs1135840	tolerated	1	benign	0	46
42126611delC	frameshift_variant		-	-	-	-	-
42126676G>A	synonymous_variant	rs150445731	-	-	-	-	-
42126749C>T	missense_variant	rs267608319	deleterious	0.04	benign	0.255	-
42126763G>T	intron_variant	rs548264542	-	-	-	-	-
42126798A>G	intron_variant	rs1243469456	-	-	-	-	-
42126914C>G	missense_variant	rs28371733	tolerated	0.14	probably damaging	0.984	-
42126938C>T	missense_variant	rs769157652	tolerated	0.22	benign	0.063	-

42126940delC	frameshift_variant		-	-	-	-	-
42126944_42126945insT	frameshift_variant		-	-	-	-	-
42126944C>T	missense_variant	rs78762568	tolerated	0.22	benign	0	-
42126963C>T	synonymous_variant	rs28371732	-	-	-	-	-
42127001G>A	intron_variant	rs4987144	-	-	-	-	25
42127019G>T	intron_variant	rs1008147170	-	-	-	-	-
42127173C>T	intron_variant	rs1249410876	-	-	-	-	-
42127207C>T	intron_variant	rs28371730	-	-	-	-	24
42127209T>C	intron_variant	rs2004511	-	-	-	-	36
42127356G>T	intron_variant	rs28371729	-	-	-	-	4
42127385C>A	intron_variant	rs891742278	-	-	-	-	-
42127398A>G	intron_variant	rs28578778	-	-	-	-	-
42127407T>G	intron_variant	rs1985842	-	-	-	-	-
42127458G>A	missense_variant	rs146271511	tolerated	0.13	possibly damaging	0.695	-
42127526_42127527insT	frameshift_variant		-	-	-	-	-
42127526C>T	missense_variant	rs1058172	deleterious	0.01	probably damaging	0.999	-
42127556T>C	missense_variant	rs202102799	deleterious	0	probably damaging	0.993	-
42127565T>C	missense_variant	rs61736517	tolerated	1	benign	0	-
42127611C>T	missense_variant	rs78209835	deleterious	0.05	benign	0	-
42127631C>G	missense_variant	rs141009491	deleterious	0	probably damaging	0.971	-
42127634C>A	missense_variant	rs3915951	deleterious	0.03	benign	0.037	-
42127644C>T	intron_variant	rs376751897	-	-	-	-	-
42127677_42127687delGTCCGGCCCTG	intron_variant	rs1237764592	-	-	-	-	-
42127694_42127695delCT	intron_variant		-	-	-	-	-
42127698T>C	intron_variant		-	-	-	-	-
42127700G>T	intron_variant		-	-	-	-	-
42127707A>G	intron_variant		-	-	-	-	-
42127718G>A	intron_variant	rs1414906948	-	-	-	-	-

Supplementary Data File S5.1 continues on next page.

Supplementary Data File S5.1: ***Continued***

Variant (NC_000022.11:g.)	Consequence	rs-number	Sift prediction	Sift score	Polyphen prediction	Polyphen score	Number of associated eQTLs according to GTEx
42127721C>T	intron_variant		-	-	-	-	-
42127734T>G	intron_variant		-	-	-	-	-
42127740T>C	intron_variant		-	-	-	-	-
42127743A>C	intron_variant		-	-	-	-	-
42127753T>G	intron_variant		-	-	-	-	-
42127755G>A	intron_variant		-	-	-	-	-
42127761C>T	intron_variant	rs267608291	-	-	-	-	-
42127774A>T	intron_variant	rs1423323203	-	-	-	-	-
42127778T>C	intron_variant		-	-	-	-	-
42127779T>C	intron_variant		-	-	-	-	-
42127782T>G	intron_variant		-	-	-	-	-
42127791G>A	intron_variant	rs536709258	-	-	-	-	-
42127792C>G	intron_variant	rs765006570	-	-	-	-	-
42127803C>T	intron_variant	rs28371725	-	-	-	-	5
42127811G>A	intron_variant	rs143276168	-	-	-	-	-
42127813C>T	intron_variant		-	-	-	-	-
42127820A>C	intron_variant		-	-	-	-	-
42127821C>T	intron_variant	rs1014423327	-	-	-	-	-
42127824T>G	intron_variant		-	-	-	-	-
42127825T>G	intron_variant		-	-	-	-	-
42127826T>C	intron_variant		-	-	-	-	-
42127832A>C	intron_variant	rs751061716	-	-	-	-	-
42127852C>T	synonymous_variant	rs79292917	-	-	-	-	-
42127853G>A	missense_variant	rs140513104	deleterious	0	possibly damaging	0.558	-

42127855A>G	synonymous_variant	rs28371724	-	-	-	-	-
42127856T>G	missense_variant	rs5030867	deleterious	0	probably damaging	1	-
42127940_42127941insA	frameshift_variant		-	-	-	-	-
42127941G>A	missense_variant	rs16947	tolerated	0.21	benign	0.062	25
42127973T>C	missense_variant	rs1135829	tolerated	0.08	benign	0.017	-
42127996G>A	intron_variant	rs371181941	-	-	-	-	-
42128042A>G	intron_variant	rs971733628	-	-	-	-	-
42128071C>G	intron_variant	rs187203531	-	-	-	-	-
42128088G>A	intron_variant	rs74516776	-	-	-	-	-
42128130C>T	intron_variant	rs28371721	-	-	-	-	-
42128136C>T	intron_variant	rs372521768	-	-	-	-	-
42128176_42128178delTCT	inframe_deletion	rs5030656	-	-	-	-	-
42128176delT	frameshift_variant	rs28371720	-	-	-	-	-
42128216G>T	synonymous_variant	rs28371718	-	-	-	-	-
42128218delG	frameshift_variant	rs72549352	-	-	-	-	-
42128242delT	frameshift_variant	rs35742686	-	-	-	-	-
42128251_42128254delTTAG	frameshift_variant	rs72549353	-	-	-	-	-
42128308C>A	missense_variant	rs28371717	tolerated	0.41	benign	0.097	-
42128321A>G	synonymous_variant	rs17002852	-	-	-	-	-
42128327G>T	synonymous_variant	rs1462273327	-	-	-	-	-
42128438C>T	intron_variant	rs79650744	-	-	-	-	-
42128500C>T	intron_variant	rs267608300	-	-	-	-	-
42128519G>A	intron_variant	rs878981790	-	-	-	-	-
42128631C>T	intron_variant	rs1468364907	-	-	-	-	-
42128693_42128694insC	intron_variant		-	-	-	-	-
42128694T>C	intron_variant	rs2267447	-	-	-	-	37
42128736C>T	intron_variant	rs761521669	-	-	-	-	-
42128741G>A	intron_variant	rs113889384	-	-	-	-	-

Supplementary Data File S5.1 continues on next page.

Supplementary Data File S5.1: ***Continued***

Variant (NC_000022.11:g.)	Consequence	rs-number	Sift prediction	Sift score	Polyphen prediction	Polyphen score	Number of associated eQTLs according to GTEx
42128793A>G	synonymous_variant	rs28371713	-	-	-	-	-
42128815C>T	missense_variant	rs5030866	tolerated	0.53	benign	0.189	-
42128858C>G	missense_variant		tolerated	0.3	benign	0	-
42128922A>G	synonymous_variant	rs111606937	-	-	-	-	-
42128945C>T	splice_acceptor_variant	rs3892097	-	-	-	-	31
42128946_42128947insT	splice_region_variant		-	-	-	-	-
42129084delA	frameshift_variant	rs5030655	-	-	-	-	-
42129087G>C	missense_variant	rs78482768	tolerated	0.88	benign	0.003	-
42129130C>G	synonymous_variant	rs1058164	-	-	-	-	46
42129174C>A	missense_variant	rs1135823	tolerated	0.39	possibly damaging	0.52	-
42129178G>A	synonymous_variant	rs61736507	-	-	-	-	-
42129180A>T	missense_variant	rs1135822	tolerated	1	benign	0	-
42129184C>T	splice_region_variant	rs267608304	-	-	-	-	-
42129225A>C	intron_variant	rs370267861	-	-	-	-	-
42129228A>G	intron_variant	rs373668214	-	-	-	-	-
42129303T>G	intron_variant	rs142302759	-	-	-	-	-
42129319_42129320delTT	intron_variant	rs1278131170	-	-	-	-	-
42129351A>G	intron_variant	rs376056664	-	-	-	-	-
42129388G>A	intron_variant	rs557722765	-	-	-	-	-
42129414G>C	intron_variant	rs143170489	-	-	-	-	-
42129436A>G	intron_variant	rs184086520	-	-	-	-	-
42129450delG	intron_variant		-	-	-	-	-
42129461G>A	intron_variant		-	-	-	-	-
42129498C>T	intron_variant	rs189736703	-	-	-	-	-

42129545G>A	intron_variant	rs267608277	-	-	-	-	-
42129581T>G	intron_variant	rs1170216892	-	-	-	-	-
42129587C>T	intron_variant	rs1183013541	-	-	-	-	-
42129614C>G	intron_variant	rs180847475	-	-	-	-	-
42129623C>T	intron_variant	rs1081004	-	-	-	-	3
42129643G>C	intron_variant	rs186133763	-	-	-	-	-
42129722C>T	intron_variant	rs368389952	-	-	-	-	-
42129726A>C	intron_variant	rs78854695	-	-	-	-	3
42129731T>C	splice_region_variant	rs267608289	-	-	-	-	-
42129754G>A	synonymous_variant	rs1081003	-	-	-	-	20
42129770G>A	missense_variant	rs28371706	tolerated	0.21	benign	0	-
42129793G>A	synonymous_variant	rs200269944	-	-	-	-	-
42129795_42129796insC	frameshift_variant		-	-	-	-	-
42129796G>C	synonymous_variant	rs28371705	-	-	-	-	30
42129809T>C	missense_variant	rs28371704	tolerated	1	benign	0	30
42129819G>T	missense_variant	rs28371703	deleterious	0.02	possibly damaging	0.927	30
42129950A>C	intron_variant	rs28371702	-	-	-	-	46
42129950delA	intron_variant		-	-	-	-	-
42130047G>C	intron_variant	rs28371701	-	-	-	-	44
42130180C>T	intron_variant	rs896350438	-	-	-	-	-
42130343T>G	intron_variant	rs867984132	-	-	-	-	-
42130469C>T	intron_variant	rs35970455	-	-	-	-	-
42130475T>C	intron_variant	rs34291018	-	-	-	-	-
42130482C>A	intron_variant	rs28371699	-	-	-	-	45
42130522G>A	intron_variant	rs29001678	-	-	-	-	5
42130538C>A	intron_variant	rs1255242741	-	-	-	-	-
42130547T>C	intron_variant		-	-	-	-	-
42130552G>A	intron_variant		-	-	-	-	-

Supplementary Data File S5.1 continues on next page.

Supplementary Data File S5.1: ***Continued***

Variant (NC_000022.11:g.)	Consequence	rs-number	Sift prediction	Sift score	Polyphen prediction	Polyphen score	Number of associated eQTLs according to GTEx
42130559T>G	intron_variant	rs28695233	-	-	-	-	-
42130560C>G	intron_variant	rs29001518	-	-	-	-	-
42130565A>G	intron_variant	rs1080998	-	-	-	-	-
42130569G>C	intron_variant	rs1080997	-	-	-	-	-
42130571G>T	intron_variant	rs1080996	-	-	-	-	-
42130578C>G	intron_variant	rs1080995	-	-	-	-	-
42130594_42130595insT	intron_variant		-	-	-	-	-
42130655_42130656insA	frameshift_variant	rs774671100	-	-	-	-	-
42130657G>A	synonymous_variant	rs757862366	-	-	-	-	-
42130667C>T	missense_variant	rs118203758	deleterious	0.02	probably damaging	0.953	-
42130692G>A	missense_variant	rs1065852	deleterious	0.02	possibly damaging	0.687	37
42130710G>A	missense_variant	rs138100349	tolerated	0.07	benign	0.021	-
42130761_42130762insT	frameshift_variant		-	-	-	-	-
42130761C>T	missense_variant	rs769258	tolerated	0.14	benign	0.024	2
42130773C>T	missense_variant	rs72549358	tolerated	0.22	benign	0	-
42130929T>C	upstream_gene_variant	rs267608272	-	-	-	-	-
42131000G>A	upstream_gene_variant	rs912915978	-	-	-	-	-
42131062T>A	upstream_gene_variant	rs1346206732	-	-	-	-	-
42131114A>G	upstream_gene_variant	rs572914357	-	-	-	-	-
42131156C>T	upstream_gene_variant	rs1080992	-	-	-	-	-
42131222G>A	upstream_gene_variant	rs566383351	-	-	-	-	-
42131379C>T	upstream_gene_variant	rs769257	-	-	-	-	-
42131469C>T	upstream_gene_variant	rs28633410	-	-	-	-	-
42131493C>T	upstream_gene_variant	rs958348536	-	-	-	-	-

42131531G>A	upstream_gene_variant	rs28624811	-	-	-	-	-
42131546_42131547delTC	upstream_gene_variant	rs536645539	-	-	-	-	-
42131610G>C	upstream_gene_variant	rs930275829	-	-	-	-	-
42131631A>T	upstream_gene_variant	rs113127784	-	-	-	-	-
42131681C>T	upstream_gene_variant	rs566307819	-	-	-	-	-
42131775C>T	upstream_gene_variant	rs567431353	-	-	-	-	-
42131791C>T	upstream_gene_variant	rs1080989	-	-	-	-	37
42131894delT	upstream_gene_variant	rs375413467	-	-	-	-	-
42131908C>T	upstream_gene_variant	rs1250789569	-	-	-	-	-
42131914_42131915delCT	upstream_gene_variant		-	-	-	-	-
42131915T>G	upstream_gene_variant	rs62625688	-	-	-	-	-
42131927C>A	upstream_gene_variant	rs528127834	-	-	-	-	-
42131990delC	upstream_gene_variant	rs757622584	-	-	-	-	-
42132026T>C	upstream_gene_variant	rs28735595	-	-	-	-	-
42132027_42132028insC	upstream_gene_variant	rs1080988	-	-	-	-	-
42132045_42132049delTTTTT	upstream_gene_variant	rs267608321	-	-	-	-	-
42132046_42132049delTTTT	upstream_gene_variant	rs267608321	-	-	-	-	-
42132047_42132049delTTT	upstream_gene_variant	rs267608321	-	-	-	-	-
42132048_42132049delTT	upstream_gene_variant	rs267608321	-	-	-	-	-
42132049_42132050insT	upstream_gene_variant	rs267608321	-	-	-	-	-
42132049_42132050insTT	upstream_gene_variant	rs267608321	-	-	-	-	-
42132049_42132050insTTT	upstream_gene_variant	rs267608321	-	-	-	-	-
42132049_42132050insTTTT	upstream_gene_variant	rs267608321	-	-	-	-	-
42132049delT	upstream_gene_variant	rs267608321	-	-	-	-	-
42132138C>G	upstream_gene_variant	rs757984982	-	-	-	-	-
42132217G>A	upstream_gene_variant	rs28588594	-	-	-	-	37
42132375G>C	upstream_gene_variant	rs1080985	-	-	-	-	19
42132377delG	upstream_gene_variant		-	-	-	-	-

Supplementary Data File S5.1 continues on next page.

Supplementary Data File S5.1: ***Continued***

Variant (NC_000022.11:g.)	Consequence	rs-number	Sift prediction	Sift score	Polyphen prediction	Polyphen score	Number of associated eQTLs according to GTEx
42132411C>T	upstream_gene_variant		-	-	-	-	-
42132561_42132562insT	upstream_gene_variant		-	-	-	-	-
42132561C>T	upstream_gene_variant	rs1080983	-	-	-	-	23
42132577G>C	upstream_gene_variant	rs1374920910	-	-	-	-	-
42132589C>T	upstream_gene_variant	rs936292274	-	-	-	-	-
42132590G>A	upstream_gene_variant	rs1054718426	-	-	-	-	-
Del							-
Dup							-