ARTICLE

Cross-Ancestry Genome-Wide Association Study Defines the Extended *CYP2D6* Locus as the Principal Genetic Determinant of Endoxifen Plasma Concentrations

Chiea Chuen Khor[1,2,3,†], Stefan Winter[4,5,†], Natalia Sutiman[6], Thomas E. Mürdter[4,5], Sylvia Chen[6], Joanne Siok Liu Lim[6], Zheng Li[1], Jingmei Li[1], Kar Seng Sim[1], Boian Ganchev[4,5], Diana Eccles[7,8], Bryony Eccles[7,8], William Tapper[7,8], Nathalie K. Zgheib[9], Arafat Tfayli[10], Raymond Chee Hui Ng[11], Yoon Sim Yap[11], Elaine Lim[11], Mabel Wong[11], Nan Soon Wong[12], Peter Cher Siang Ang[12], Rebecca Dent[11], Roman Tremmel[4,5], Kathrin Klein[4,5], Elke Schaeffeler[4,5,13], Yitian Zhou[14], Volker M. Lauschke[4,5,15], Michel Eichelbaum[4,5], Matthias Schwab[4,13,16,17,18], Hiltrud B. Brauch[4,5,13,18], Balram Chowbay[3,6,19,*,†] and Werner Schroth[4,5,*,†]

The therapeutic efficacy of tamoxifen is predominantly mediated by its active metabolites 4-hydroxy-tamoxifen and endoxifen, whose formation is catalyzed by the polymorphic cytochrome P450 2D6 (CYP2D6). Yet, known *CYP2D6* polymorphisms only partially determine metabolite concentrations *in vivo*. We performed the first cross-ancestry genome-wide association study with well-characterized patients of European, Middle-Eastern, and Asian descent (n = 497) to identify genetic factors impacting active and parent metabolite formation. Genome-wide significant variants were functionally evaluated in an independent liver cohort (n = 149) and *in silico*. Metabolite prediction models were validated in two independent European breast cancer cohorts (n = 287, n = 189). Within a single 1-megabase (Mb) region of chromosome 22q13 encompassing the *CYP2D6* gene, 589 variants were significantly associated with tamoxifen metabolite concentrations, particularly endoxifen and metabolic ratio (MR) endoxifen/*N*-desmethyltamoxifen (minimal P = 5.4E−35 and 2.5E−65, respectively). Previously suggested other loci were not confirmed. Functional analyses revealed 66% of associated, mostly intergenic variants to be significantly correlated with hepatic CYP2D6 activity or expression (ρ = 0.35 to −0.52), and six hotspot regions in the extended 22q13 locus impacting gene regulatory function. Machine learning models based on hotspot variants (n = 12) plus CYP2D6 activity score (AS) increased the explained variability (~9%) compared with AS alone, explaining up to 49% (median R^2) and 72% of the variability in endoxifen and MR endoxifen/*N*-desmethyltamoxifen, respectively. Our findings suggest that the extended *CYP2D6* locus at 22q13 is the principal genetic determinant of endoxifen plasma concentration. Long-distance haplotypes connecting *CYP2D6* with adjacent regulatory sites and nongenetic factors may account for the unexplained portion of variability.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ The therapeutic efficacy of tamoxifen likely depends on bioactivation to active metabolites; however, interindividual differences in plasma concentrations of the major active metabolite endoxifen are only partially explained by known *CYP2D6* polymorphisms.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ Are there genetic factors in addition to *CYP2D6* that impact active and parent metabolite formation and to what extent do they improve the prediction of variable endoxifen concentrations?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ Endoxifen formation largely depends on a *CYP2D6*-encompassing extended chr22q13 locus with intergenic variants linked to CYP2D6 function in liver and to *in silico*–predicted regulatory function. Models that combine CYP2D6 activity score and surrounding variants enhance the genome-based prediction of active tamoxifen metabolite levels.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ Long-range genetic interactions in the 22q13 region derived from haplotype data may improve the prediction of endoxifen variability through comprehensive assessment of CYP2D6 activity, potentially leading to reevaluation of its use as a biomarker of tamoxifen response.

Received October 28, 2022; accepted December 23, 2022. doi:10.1002/cpt.2846

[1]Division of Human Genetics, Genome Institute of Singapore, Singapore, Singapore; [2]Singapore Eye Research Institute, Singapore, Singapore; [3]Clinical Pharmacology, SingHealth, Singapore, Singapore; [4]Dr Margarete Fischer-Bosch Institute of Clinical Pharmacology, Stuttgart, Germany; [5]University Tübingen, Tübingen, Germany; [6]Clinical Pharmacology Laboratory, Division of Cellular and Molecular Research, National Cancer Centre, Singapore, Singapore; [7]Faculty of Medicine, Cancer Sciences Academic Unit and University of Southampton Clinical Trials Unit, University of Southampton, Southampton, UK; [8]University Hospital Southampton National Health Service Foundation Trust, Southampton, UK; [9]Department of Pharmacology and Toxicology, Faculty of Medicine, American University of Beirut, Beirut, Lebanon; [10]Hematology-Oncology Division, Department of Internal Medicine, Faculty of Medicine, American University of Beirut, Beirut, Lebanon; [11]Division of Medical Oncology, National Cancer Centre, Singapore, Singapore; [12]OncoCare Cancer Centre, Mount Elizabeth Novena Medical Centre, Singapore, Singapore; [13]Image-Guided and Functionally Instructed Tumor Therapies Cluster of Excellence (iFIT), University of Tübingen, Tübingen, Germany; [14]Department of Laboratory Medicine, Karolinska Institute, Stockholm, Sweden; [15]Department of Physiology and Pharmacology, Karolinska Institute, Stockholm, Sweden; [16]Department of Clinical Pharmacology, University of Tübingen, Tübingen, Germany; [17]Department of Biochemistry and Pharmacy, University of Tübingen, Tübingen, Germany; [18]German Cancer Consortium (DKTK), German Cancer Research Center, Partner Site Tübingen, Tübingen, Germany; [19]Centre for Clinician-Scientist Development, Duke–National University of Singapore Medical School, Singapore, Singapore. *Correspondence: Balram Chowbay (ctebal@nccs.com.sg), Werner Schroth (werner.schroth@ikp-stuttgart.de)

[†]Contributed equally.

Adjuvant therapy of early estrogen receptor (ER)–positive breast cancer with the selective ER modulator tamoxifen is a well-established modality for the reduction of hormone-sensitive breast cancer recurrence and mortality. Clinical trial evidence showed that after 10 years, 88% of adjuvantly treated patients were still alive and 77% were free of recurrences; however, long-term clinical failure occurs in up to one-third of treated patients.[1] Of several mechanisms that contribute to clinical nonresponse, impaired bioactivation of tamoxifen to its active metabolites (Z)-4-hydroxytamoxifen and (Z)-*N*-desmethyl-4-hydroxy-tamoxifen (endoxifen) may constitute an intrinsic feature referred to as metabolic resistance.[2,3] Endoxifen is considered the major therapeutic metabolite based on its 5 to 7 times higher plasma concentrations compared with (Z)-4-hydroxytamoxifen[4,5] and blocking of ER-mediated tumor growth.[6] However, the abundant but less potent parent drug tamoxifen and *N*-desmethyl tamoxifen as well as other metabolites also exert some inhibitory effects at the ER.[6,7]

The formation of endoxifen is catalyzed by cytochrome P450 (CYP) oxidases, of which the CYP2D6 enzyme is most prominent.[8] CYP2D6 shows extensive interindividual functional variability resulting in four major phenotype groups: individuals with increased, normal, reduced and no CYP2D6 activity are categorized as ultrarapid (UM), normal (NM), intermediate (IM), and poor (PM) metabolizers, respectively. This variation is attributed to more than 100 genetic *CYP2D6* variants of which the splice acceptor polymorphism *CYP2D6*4* (rs3892097, frequency up to 23%) and the gene deletion *5 (frequency up to 6%) define common loss-of-function alleles in Europeans, whereas the reduced function variants *CYP2D6*10* (rs1065852) and *CYP2D6*41* (rs28371725) are most prevalent in Asian (45% frequency) and European/Middle-Eastern populations (7–20% frequency), respectively. Comprehensive genotyping of the most common variants is translated into an activity score (AS) as a semiquantitative measurement of enzyme activity in drug metabolism studies.[9,10]

Reduced endoxifen plasma concentrations were associated with clinical outcome with suggested critical therapeutic thresholds of 9–16 nM, above which increased clinical benefit may be expected.[11–13] However, there are controversies as some studies did not confirm the predictive role of endoxifen plasma concentrations or CYP2D6.[14–16] Since known *CYP2D6* alleles account for only about 10–40% of endoxifen variability[3,17] it has been suggested that additional genetic and nongenetic factors exist that contribute to differences in metabolite concentrations.[18–22] Recently, a genome-wide association study (GWAS) of 192 European patients suggested that in addition to *CYP2D6* other genomic regions may influence such differences;[23] however, no independent validation was provided. Based on long-range *CYP2D6* gene sequencing analysis and a deep neural network model, another study suggested an improved prediction of the CYP2D6-dependent N-desmethyl tamoxifen–to-endoxifen formation with a continuous enzyme activity scale.[24] To shed new light on the question whether genetic factors other than known *CYP2D6* variants contribute to the impaired formation of endoxifen and its precursor metabolites (tamoxifen, N-desmethyl tamoxifen, and (Z)-4-hydroxytamoxifen) across populations, we conducted a cross-ancestry GWAS in three ethnic populations, explored the functional effect of variants by *in silico* analyses and on CYP2D6 expression and activity in an independent liver cohort, and validated multi–single-nucleotide variant (SNV, formerly SNP) prediction models for tamoxifen metabolites in two independent European cohorts. Here, we report that a single genomic region at chromosome 22q comprising *CYP2D6* has significant influence on tamoxifen biotransformation, and that six hotspot regions including multiple variants with putative regulatory function (upstream and downstream of the coding region) may co-influence endoxifen plasma concentrations in patients with early breast cancer treated with tamoxifen.

METHODS

Breast cancer patient collections and data

Patients and specimens used in this study are summarized in the study flowchart (**Figure 1**). Our cross-ancestry GWAS was based on three cohorts: 154 Singaporean Chinese premenopausal and postmenopausal patients who were histologically diagnosed with hormone receptor (HR)–positive breast cancer and prospectively recruited at the National Cancer Centre, Singapore ("Singapore" cohort); 70 premenopausal patients with HR-positive breast cancer recruited at the American University of Beirut Medical Centre, Lebanon ("Lebanon" cohort); and 290 postmenopausal patients with HR-positive breast cancer obtained from the tamoxifen arm of a prospective, observational multicenter adjuvant endocrine treatment study ("Germany" cohort, IKP211 study; German Clinical Trial Register DRKS 00000605[5,25]). Validation of multi-SNV prediction models was based on two European patient cohorts: 287 HR-positive patients from the Prospective Study of Outcomes in Sporadic vs. Hereditary Breast Cancer (POSH) cohort[26] from the University of Southampton, UK ("UK" cohort), and available genome-wide genotype data and plasma metabolite concentrations of 189 patients from the Marie Sklodowska-Curie Memorial Cancer Center and Institute of Oncology in Warsaw, Poland, downloaded

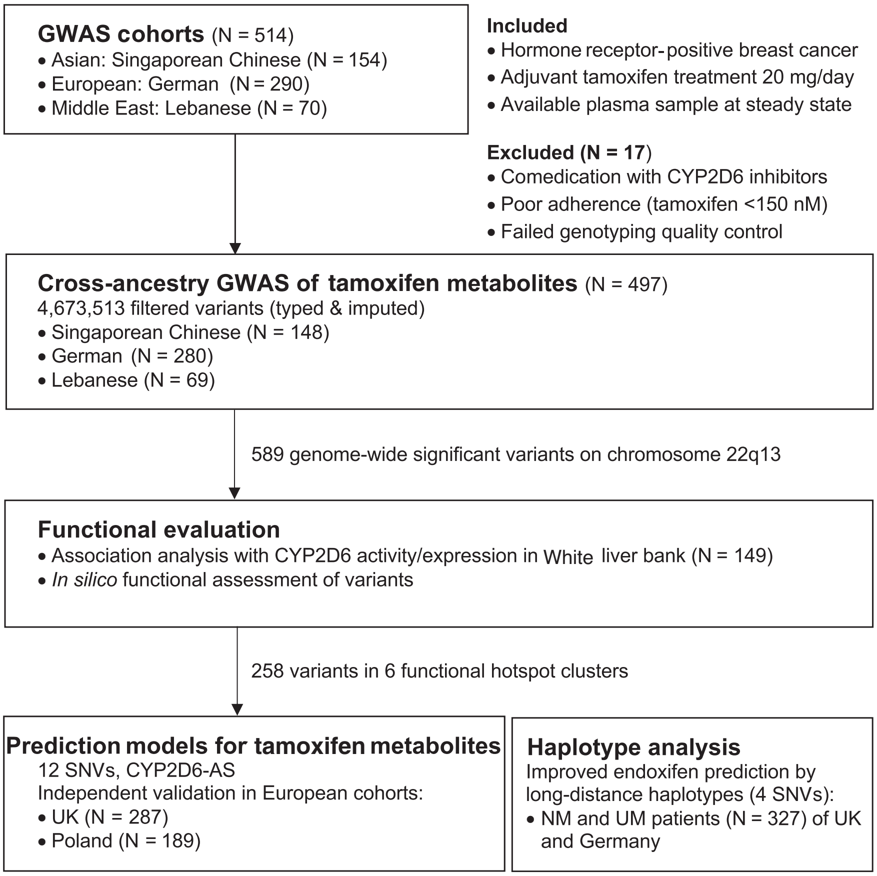

GWAS cohorts (N = 514)
• Asian: Singaporean Chinese (N = 154)
• European: German (N = 290)
• Middle East: Lebanese (N = 70)
Included
• Hormone receptor-positive breast cancer
• Adjuvant tamoxifen treatment 20 mg/day
• Available plasma sample at steady state
Excluded (N = 17)
• Comedication with CYP2D6 inhibitors
• Poor adherence (tamoxifen <150 nM)
• Failed genotyping quality control
Cross-ancestry GWAS of tamoxifen metabolites (N = 497)
4,673,513 filtered variants (typed & imputed)
• Singaporean Chinese (N = 148)
• German (N = 280)
• Lebanese (N = 69)
589 genome-wide significant variants on chromosome 22q13
Functional evaluation
• Association analysis with CYP2D6 activity/expression in White liver bank (N = 149)
• In silico functional assessment of variants
258 variants in 6 functional hotspot clusters
Prediction models for tamoxifen metabolites
12 SNVs, CYP2D6-AS
Independent validation in European cohorts:
• UK (N = 287)
• Poland (N = 189)
Haplotype analysis
Improved endoxifen prediction by long-distance haplotypes (4 SNVs):
• NM and UM patients (N = 327) of UK and Germany

Figure 1 Study flowchart diagram. CYP2D6, cytochrome P450 2D6; CYP2D6-AS, cytochrome P450 2D6 activity score; GWAS, genome-wide association study; NM, normal metabolizer; SNV, single-nucleotide variant; UM, ultrarapid metabolizer.

from NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/, accession number: GSE129162; "Poland" cohort).[23] All patients had received 20-mg tamoxifen daily for at least 8 weeks prior to blood sampling (steady state). The CYP2D6-AS was inferred from previous genotyping of major alleles including **2, *3, *4, *5, *6, *7, *9, *10, *35, *41* and gene duplication which were translated into ASs categorizing four metabolizer phenotypes: 0 (poor), 0.25 to 1 (intermediate), 1.25 to 2.25 (normal), and ≥2.5 (ultrarapid) as previously defined.[10] Study inclusion and exclusion criteria for all five cohorts have been previously described.[5,23,26–28]

Informed consent and ethics

The study has been carried out in accordance with the provisions of the Declaration of Helsinki of 1975. Ethics approval was obtained from the ethics committee of the National Cancer Centre Singapore (Singapore), the American University of Beirut (AUB) institutional ethics review board (Lebanon), the Medical Faculty of the University of Tübingen and the local ethics committees of all participating centres in Germany (IKP211 study), and South and West MultiCentre Research Ethics committee (POSH, UK). Informed patient consent was obtained from all participants as required by institutional review boards and research ethics committees. All patient data were deidentified prior to inclusion in this study. Analyses of liver specimens were approved by the ethics committees of the medical faculties of the Charité, Humboldt University, Germany, and of the University of Tuebingen, Germany as described.[29]

Measurements of steady-state blood concentrations of tamoxifen and metabolites

Whole-blood samples (3 mL) were drawn from tamoxifen-treated patients of the GWAS and UK cohorts after at least 8 weeks of tamoxifen therapy (20 mg/day). Plasma was obtained by centrifugation under light protection within 30 minutes of venipuncture and stored at −80°C until analysis. Steady-state plasma concentrations of tamoxifen, N-desmethyl tamoxifen, and (Z)-isomers of the active metabolites, (Z)-4-hydroxytamoxifen and endoxifen, were quantified by liquid chromatography tandem mass spectrometry in the multiple reaction monitoring mode performed on an Agilent 1290 Series Rapid Resolution LC System coupled to a 6,460 triple quadrupole mass spectrometer (Agilent Technologies, Waldbronn, Germany) as previously described.[5]

Genome-wide genotyping

Genome-wide genotyping of patients was performed using the Illumina Infinium OmniExpress 12/24 beadchips platform (Illumina, Singapore), following manufacturer's instructions (http://www.illumina.com). Genome-wide genotyping data of the Polish cohort were downloaded from NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/, accession number: GSE129162, platform: Illumina HumanOmni2.5-8 BeadChip). Quality control, imputation analysis, variant definitions, and population stratification analysis are described in **Materials and Methods S1**.

Genome-wide association analyses in individual studies and cross-ancestry GWAS

GWAS was performed for six pharmacokinetic end points including endoxifen, (Z)-4-hydroxytamoxifen, N-desmethyl tamoxifen, and tamoxifen as well as metabolic ratios (MRs) endoxifen/N-desmethyl tamoxifen and (Z)-4-hydroxytamoxifen/tamoxifen based on typed and imputed variants in 148 Singaporean Chinese, 280 German, and 69 Lebanese patients. Patients that had received strong CYP2D6 inhibitors or had

Table 1 Characteristics and tamoxifen metabolite plasma concentrations of patients used in cross-ancestry GWAS and for prediction model validations

Patient characteristics and plasma concentrations	*Cross-ancestry GWAS cohorts*			*Model validation cohorts*	
	Singapore (*n* = 148)	Germany (*n* = 280)	Lebanon (*n* = 69)	UK (*n* = 287)	Poland (*n* = 189)
Age, median (range), y	49 (31–70)	64 (45–82)	43 (24–51)	38 (22–41)	54 (25–90)
Weight, median (range), kg	56 (39–92)	71 (45–144)	69 (41–110)	65 (44–124)	NA
Height, median (range), cm	156 (134–172)	163 (150–180)	162 (146–175)	165 (132–183)	NA
BMI, median (range), kg/m[a]	23 (14–38)	26 (18–41)	25 (15–41)	24.2 (16.8–45.4)	NA
Menopausal status, *N* (%)					
Premenopausal	120 (81.1)	10 (3.6)	69 (100)	287 (100)	83 (28.3)
Postmenopausal	28 (18.9)	268 (95.7)	0 (0)	0 (0)	124 (42.3)
Unknown	0 (0)	2 (0.7)	0 (0)	0 (0)	86 (29.4)
Prior cancer treatment, *N* (%)					
Chemotherapy	102 (68.9)	61 (22.4)	60 (87.0)	225 (74.0)	NA
Unknown	28 (18.9)	5 (1.7)	0 (0)	0 (0)	NA
Receptor status, *N* (%)					
ER+	134 (90.5)	278 (99.3)	68 (98.6)	297 (98)	NA
PR+	132 (89.2)	239 (85.4)	55 (79.7)	166 (54.8)	NA
Metabolite C_{ss}, median (range), nM					
Tamoxifen	536.2 (164.5–1,246.8)	417.3 (150.7–2,608.8)	389.9 (161.1–795)	367.1 (155.4–1,061.2)	444.3 (152.9–1,067.7)
N-desmethyl tamoxifen	1015.4 (285.5–2,507)	716.8 (247.2–2,014)	722.1 (274.6–1,286.9)	690.7 (149.7–1,948)	625.4 (122.6–2,268.5)
(Z)-endoxifen	42.4 (5.5–142.5)	28.4 (5.8–95.4)	35.5 (7.8–88.3)	25.3 (2.4–105.9)	11.8 (1.6–48.7)[b]
(Z)-4-hydroxytamoxifen	6.9 (2.2–18.5)	5.8 (1.4–22.3)	5.7 (2.4–16.2)	5.86 (1.64–18.7)	5.8 (0.3–14.6)[a]

BMI, body mass index; C_{ss}, steady-state concentration; ER+, estrogen receptor-positive; GWAS, genome-wide association study; NA, not available; PR+, progesterone receptor-positive.

[a]Quantified as sum of (Z)-4-OH-tamoxifen + 3-OH-tamoxifen[29]. [b]Quantified as sum of endoxifen +3-OH-NDM-tamoxifen[29].

tamoxifen concentrations below 150 nM were excluded (**Figure 1**). Results of association analyses in individual cohorts were combined via inverse-variance weighted fixed-effects meta-analysis ("cross-ancestry GWAS") using R-package metafor v2.4-0[30] as described in **Materials and Methods S1**. Genome-wide significance level was defined as 5E−08.

Correlation analyses of significant variants with CYP2D6 enzyme activity and protein

Associations between genome-wide significant variants and microsomal CYP2D6 enzyme activity or protein expression were investigated in 149 subjects of a European liver cohort,[29] as described in the **Material and Methods S1**.

Functional analyses of candidate variants by bioinformatic prediction tools

Genome-wide significant variants in the chromosome 22q13 region were analyzed using five computational tools to assess functional consequences of noncoding variation as described in the **Material and Methods S1**. Hotspot clusters were defined as regions enriched in functional evidence by the presence of ≥4 consecutive variants sharing the same functional prediction and containing a variant with evidence derived from two different functional analyses, or by the site with strongest *in silico* signal plus 3–4 flanking variants.

Multi-SNV prediction models (European cohorts)

Multi-SNV models for the prediction of tamoxifen metabolites were based on the genome-wide significant variants in the hotspot clusters determined by functional analysis. Feature selection resulted in multi-SNV sets for endoxifen ($n = 8$ variants) and MR endoxifen/N-desmethyl tamoxifen ($n = 12$). These multi-SNV sets were used for the prediction of the two metabolic end points in the German cohort applying the ensemble machine learning framework implemented in R-package SuperLearner v2.0-28,[31] thereby considering six different machine learning algorithms. SuperLearner predictions were then used to determine model performance (R^2) in the validation cohorts (UK and Poland). Details are described in the **Material and Methods S1**.

Haplotype analysis of chromosome 22 variants (European cohorts)

Genotype data of four variants (*CYP2D6**2* variants rs16947 and rs1135840, the *41 defining rs28371725[32] and rs5758550, located 114 kilobases (kb) downstream of *CYP2D6* and reported to enhance *CYP2D6* promotor activities[33]) were used to test for an effect of long-distance haplotypes on the prediction of metabolite concentrations in normal metabolizer patients (CYP2D6 UM, NM/NM or NM/IM) of the combined German and UK cohort ($n = 327$) as described in the **Material and Methods S1**.

RESULTS

Cross-ancestry GWAS

Demographics, clinical characteristics, and steady-state tamoxifen metabolite plasma concentrations of all patients are given in **Table 1**. Principal component analysis confirmed that all patients of the GWAS cohorts were genetically homogenous within their own population strata (**Figure S1**). Cross-ancestry GWAS for the six pharmacokinetic end points showed no genomic inflation (**Figure S2F**), suggesting that the association results were not confounded by cryptic population substructure. In the meta-analysis we did not observe genome-wide significant associations with tamoxifen (**Figure S2A**), indicating that genetic influence on blood tamoxifen concentrations, if any, is not detectable within our study. In contrast, a total of 589 variants within an ~ 1-Mb region mapping to chromosome 22q13 (chr22: 41752944–42695148, GRCh37; **Data S1**) revealed highly significant associations for endoxifen ($P_{rs56023519}$ = 5.4E−35; **Figure 2a**) and MR endoxifen/N-desmethyl tamoxifen ($P_{rs56023519}$ = 2.5E−65), followed by N-desmethyl tamoxifen ($P_{rs5751245}$ = 4.8E−14), (Z)-4-hydroxytamoxifen ($P_{rs3021082}$ = 3.7E−09), and MR (Z)-4-hydroxytamoxifen/tamoxifen ($P_{rs56023519}$ = 3.6E−21; **Figure S2B–E**). There was a single genome-wide significant hit outside the 22q13 region for MR (Z)-4-hydroxytamoxifen/tamoxifen, with two highly linked variants on chromosome 6 mapping to the *RIPOR2* gene (minimal *P* = 2.9E−08). Due to their much lesser effect compared with the SNVs in the chromosome 22 region and lack of significance upon covariate adjustment, the two variants were not further followed in this study. Variants at chromosomes 3, 5, 7, 8, and 13 previously suggested to influence endoxifen metabolism in a study of Polish patients,[23] and a recently reported nuclear factor (NFIB) variant on chromosome 9 that regulates CYP2D6[34] did not replicate in our study (**Figures 2a and S2**). Regional association plots indicated that the identified chromosome 22 locus encompasses a region containing 25 mapped genes, including *CYP2D6* (**Figures 3a and S3A**).Next, we accounted for covariates reported to affect endoxifen plasma concentrations: when adjusting for age, weight, *CYP2C9*, *CYP2C19*, and *CYP3A5* variants, genome-wide significance was retained at the chromosome 22 locus (minimal *P* = 8E−23; **Figure 2b**). When we additionally accounted for the CYP2D6-AS that was calculated based on known functionally relevant haplotypes, a portion of chromosome 22 variants still remained genome-wide significant (minimal *P* = 4E−9; **Figure 2c**). The significance for the two chromosome 6 variants associated with MR (Z)-4-hydroxytamoxifen/tamoxifen vanished upon non-*CYP2D6* covariate adjustment (minimal *P* = 2E−04). Moreover, none of the chromosome 22 variants associated with (Z)-4-hydroxytamoxifen, (Z)-4-hydroxytamoxifen/tamoxifen, and N-desmethyl tamoxifen retained genome-wide significance upon CYP2D6-AS adjustment, a reason why we focused on endoxifen and MR endoxifen/N-desmethyl tamoxifen in subsequent analyses. A regional association plot for the cross-ancestry GWAS of endoxifen with adjustment for non-*CYP2D6* covariates and CYP2D6-AS revealed significant variants within 120 kb encompassing *CYP2D7/CYP2D6* plus a downstream region (top lead SNV rs6002629, *P* = 4E−09; **Figure 3b**). An almost identical region was associated with MR endoxifen/N-desmethyl tamoxifen (**Figure S3B**).

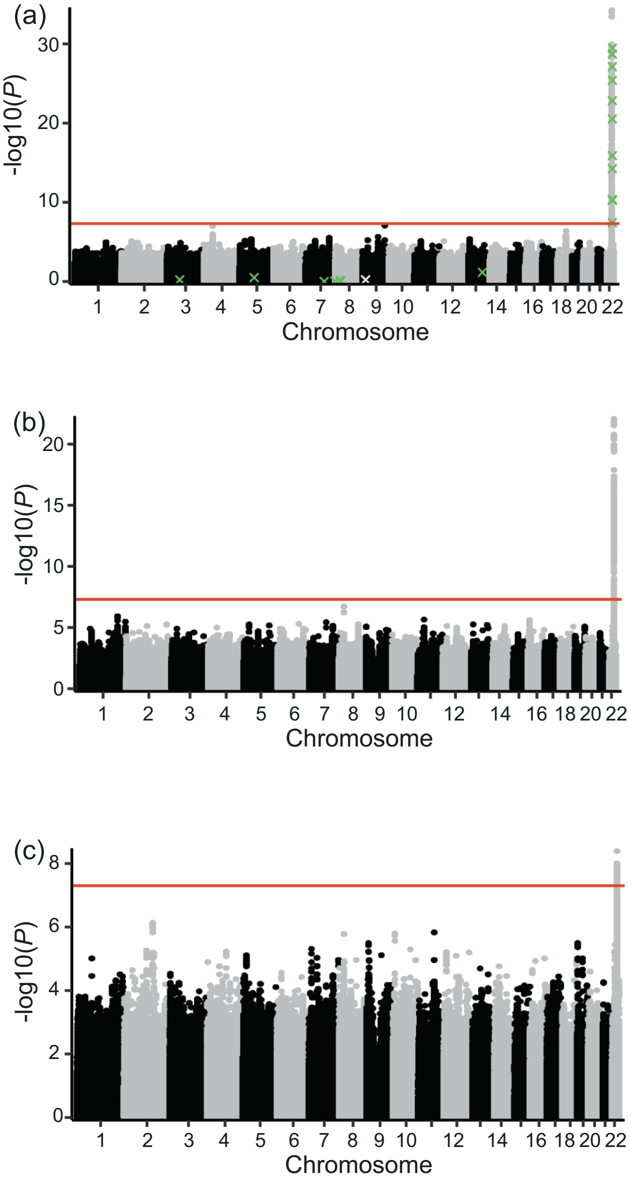


Figure 2 Manhattan plots showing association *P* values (−log10-transformed) of cross-ancestry GWAS for log (Z)-endoxifen (*n* = 497). Meta-analyses with adjustment for (**a**) top PCs (principal components) only, (**b**) PCs and covariates weight, age, as well as non-*CYP2D6* variants CYP2C9*2 and *3, CYP2C19*2 and *17, CYP3A5*3 ("non-*CYP2D6*"), and (**c**) PCs, non-*CYP2D6* covariates and CYP2D6-AS. Genome-wide significance level (5E−08) is indicated by the red line. Significant hits reported in previous GWAS for endoxifen variability[23] and an NFIB variant on chromosome 9 that regulates CYP2D6[34] are marked by green and white asterisks, respectively in A. Strong associations were observed at chromosome 22 (significant variants are listed in **Data S1**). There was no genome-wide significant association at other chromosomal regions. CYP2D6-AS, cytochrome P450 2D6 activity score; GWAS, genome-wide association study; NFIB, nuclear factor I B.

Functional assessment of the identified 22q13 variants

508 of the 589 genome-wide significant variants in our cross-ancestry GWAS, genome-wide genotype data was available for 149 subjects of a European liver tissue bank. Of these variants, 338 (66%) were significantly correlated with CYP2D6 enzyme activity or protein expression (Negative, *n* = 214: Spearman's *ρ* −0.52 to −0.18; Positive, *n* = 124: Spearman's *ρ* 0.19 to 0.35; Benjamini-Hochberg adjusted *P* ≤ 0.05; **Figures 4 and S4; Data S2**). The vast majority (90%) were located within ≈350 kb

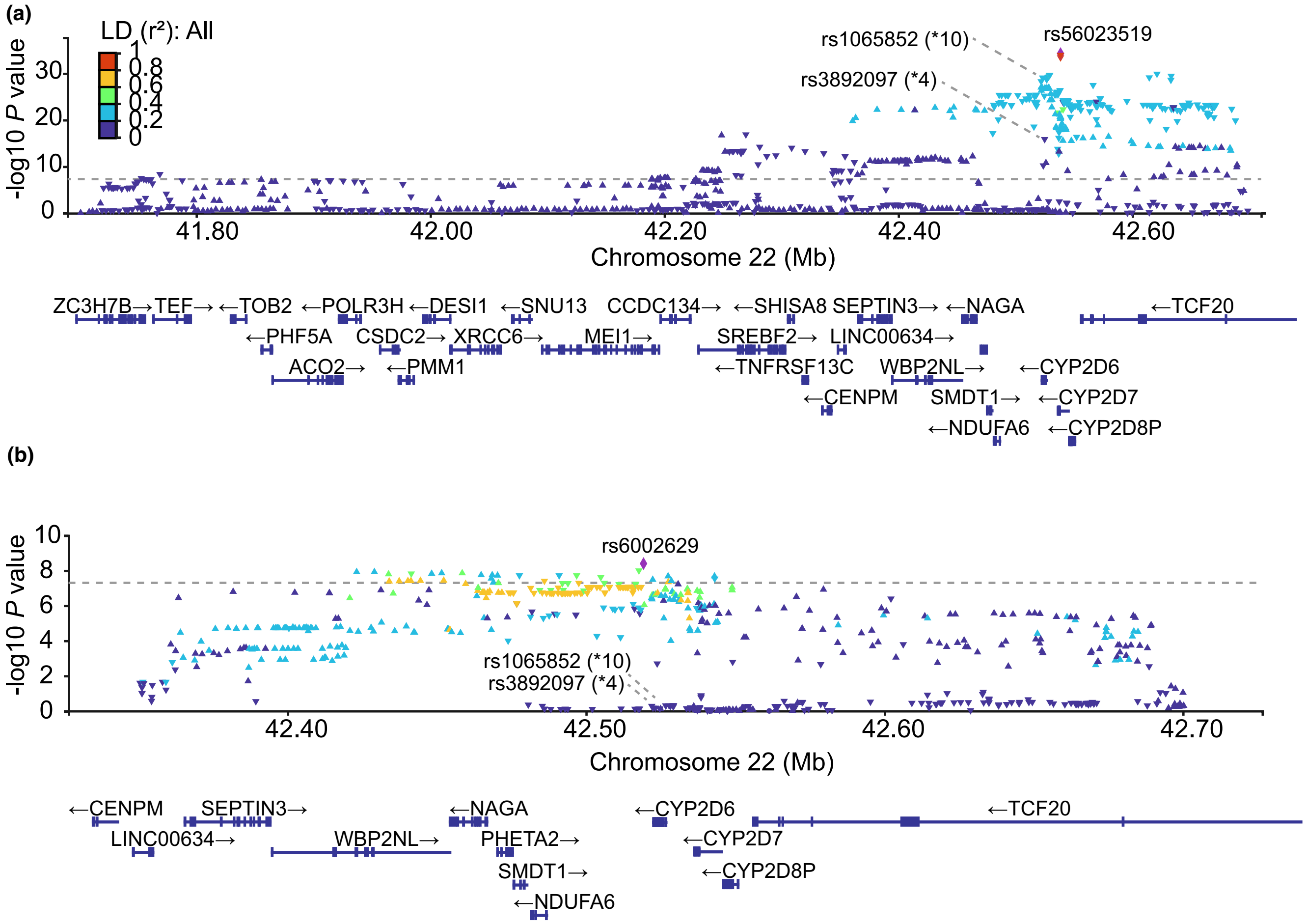


Figure 3 Regional association plots of the chromosome 22 locus for log (Z)-endoxifen. (**a**) Cross-ancestry GWAS (n = 497) with adjustment for top principal components. The top SNV rs56023519 maps to CYP2D7, 11 kilobases (kb) upstream of CYP2D6; the two major CYP2D6 variants rs3892097 (*4) and rs1065852 (*10) are shown; LD (r^2) color codes refer to pairwise comparisons with rs56023519 in all populations (ALL) of the 1000 Genomes Phase 3 data. (**b**) Cross-ancestry GWAS (n = 497) adjusted for weight, age, CYP2C9*2 and *3, CYP2C19*2 and *17, CYP3A5*3 as well as known CYP2D6 alleles represented by AS. A top SNV (rs6002629) located 5 kb downstream of CYP2D6 and linked variants (LD of $r^2 \geq 0.5$) map to a region encompassing 120 kb, pointing to a genetic component that is not captured by the known CYP2D6 haplotypes. The genome-wide significance level (5E–08) is indicated by the horizontal dashed line. Significant variants are listed in **Data S1** (sheets: logEndoxifen_pconly and logEndoxifen_covariates_AS). AS, activity score; CYP2D6, cytochrome P450 2D6; LD, linkage disequilibrium; Mb, megabase; SNV, single-nucleotide variant; gene names are listed according to HUGO Gene Nomenclature Committee name definitions.

upstream and downstream of *CYP2D6*. Variants in the ≈170 kb upstream region showed strongest correlations, had low to high LD to the major European *CYP2D6*4* rs3892097 variant (r^2 = 0–0.92, **Data S2**), and were mainly associated with a deleterious effect on CYP2D6 activity (**Figure S4**). Two variants that were previously reported to either influence CYP2D6 messenger RNA expression and activity (rs5751247)[35] or to enhance messenger RNA expression (rs5758550)[33] were confirmed for their CYP2D6 correlation (ρ = −0.45, Benjamini-Hochberg adjusted P = 5.5E−07; ρ = 0.21, Benjamini-Hochberg adjusted P = 0.02, respectively).

To predict the probability of functional impacts of variants in noncoding regions, five *in silico* algorithms were applied to all 589 genome-wide significant variants in the 22q13 region. Overall, we identified six hotspot regions with putative regulatory relevance, comprising 258 of the 589 variants (**Figure 4** and **Data S2**; positions relative to the *CYP2D6* transcription start site (TSS)). Two clusters were upstream of the gene body (cluster 1 from −168 to −137 kb relative to the TSS; cluster 2 from −92 to −82 kb), one encompassed the *CYP2D6* gene and its immediate upstream interval (cluster 3 from −42 to +10 kb) and three were localized downstream of the gene (cluster 4, 5, and cluster 6 from +38 to 42 kb, from +50 to 63 kb and from +173 to 183 kb, respectively). These findings indicate that regulatory variation is found in the immediate proximity of *CYP2D6* and, moreover, spans over genomic intervals up to 200 kb away from the TSS, thus highlighting regions for further high-resolution functional profiling.

In an attempt to clinically prioritize relevant regions, we performed an association analysis of 22q13 variants with tamoxifen adverse drug reactions (hot flashes, depression, endometrial carcinoma) in patients of the German GWAS cohort with available data. Since the significance of few variants in hotspot cluster regions 3 and 5 vanished after correction for multiple testing (not shown), this clinical phenotype was not further considered.

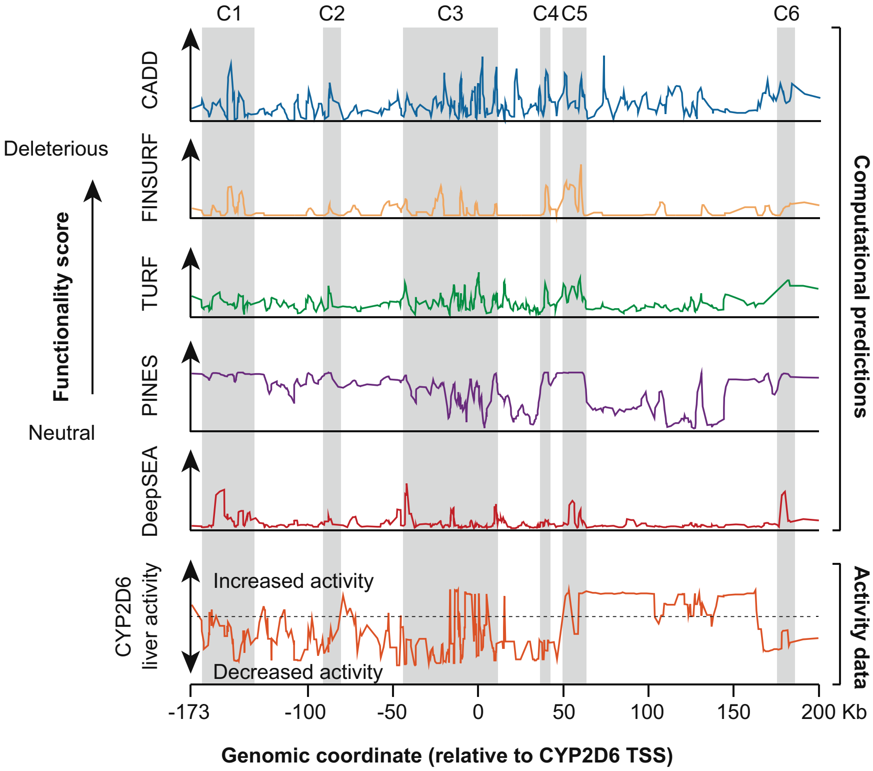


Figure 4 Functional mapping of the chromosome 22q13 region. The 589 genome-wide significant variants associated with tamoxifen metabolite concentrations were annotated with five computational prediction tools (upper tracks) and CYP2D6 activity or protein expression from 149 liver specimens (bottom track): CAAD and FINSURF predict deleteriousness and pathogenicity, respectively; TURF predicts functional probabilities for regulatory variants in liver; PINES predicts pathogenic effects of noncoding variants based on liver-specific epigenetic annotations, deepSEA predicts decreased or increased regulatory activity; CYP2D6 liver activity refer to positive (up) or negative (down) Spearman's ρ correlation coefficients (smoothened trendline based on Microsoft Excel's moving average function with period 4). Six clusters (C1–C6) of enriched functional evidence were defined by the presence of ≥4 consecutive variants sharing the same functional prediction and containing a variant with evidence derived from two different functional analyses, or by the site with strongest *in silico* signal plus 3–4 flanking variants (**Data S2**): two clusters were upstream of the gene body (cluster 1 from −168 to −137 kb (kilobases) relative to the TSS; cluster 2 from −92 to −82 kb), one encompassed the *CYP2D6* gene and its immediate upstream interval (cluster 3 from −42 to 10 kb) and three were localized downstream of the gene (cluster 4, 5, and cluster 6 from +38 to 42 kb, from +50 to 63 kb and from +173 to 183 kb, respectively). C, cluster; CADD, Combined Annotation Dependent Depletion; CYP2D6, cytochrome P450 2D6; deepSEA, deep learning-based algorithmic framework for predicting the chromatin effects of sequence alterations with single nucleotide sensitivity; FINSURF, machine-learning approach to predict the functional impact of non-coding variants in regulatory regions; PINES, Phenotype-Informed Noncoding Element Scoring; TSS, transcription start site, TURF, Tissue-specific Unified Regulatory Features.

Single-marker and multi-SNV models for endoxifen and MR endoxifen/N-desmethyl tamoxifen prediction

Single top candidates (selected by minimum GWAS *P* value of each hotspot cluster) explained only a fraction of the variability for both endoxifen and MR endoxifen/N-desmethyl tamoxifen metabolite end points (median R^2 (%): Endoxifen: 2–23%, MR: 4–38%), compared with CYP2D6-AS (Endoxifen: 30–40%, MR: 49–63%). Therefore, we applied ensemble machine learning to investigate to what extent the endoxifen and MR endoxifen/N-desmethyl tamoxifen variability can be explained by genetic factors and whether the prediction by classical CYP2D6-AS can be improved by consideration of *CYP2D6*-neighboring meta-analysis hits. Feature selection based on the 258 variants in the six hotspot clusters revealed 12 SNVs (including 5 *CYP2D6* variants; **Table S1**), of which 8 SNVs were used for endoxifen and all 12 SNVs for MR endoxifen/N-desmethyl tamoxifen prediction. Ensemble machine learning revealed a lower performance in explaining endoxifen variability for the multi-SNV model (median R^2 (%): 24.4–31.8%, model 2, **Table 2**) as compared with the reference CYP2D6-AS (median R^2 (%): 32.9–40.5%, model 1). Yet, the explained variability increased on average to 35–49.3% when CYP2D6-AS was added to the SNV set (model 3). The explained variability of MR endoxifen/N-desmethyl tamoxifen was similarly improved by 6–9% when the multi-SNV set was combined with CYP2D6-AS (median R^2 (%): 60.8–72.2%, model 3) compared with CYP2D6-AS alone (model 1). Inclusion of the other important pharmacogene variant alleles *CYP2C9*2* and **3*, *CYP2C19*2* and **17*, and *CYP3A4*22*[22] did only marginally enhance the average model performance (1.7% and 0.4% increase in median R^2 for endoxifen and MR endoxifen/N-desmethyl tamoxifen, respectively; data not shown). Moreover, the explained variability did not relevantly change when all variants in the 1-Mb region were considered instead of limiting feature selection to the 258 cluster variants.

Table 2 Percentage of variance (R^2, %) in endoxifen concentrations and MR endoxifen/N-desmethyl tamoxifen explained by prediction models in European patients with breast cancer

		Endoxifen			MR endoxifen/N-desmethyl tamoxifen[a]		
		Training	Validation		Training	Validation	
Model[b]	Model based on	Germany[c]	UK[d]	Poland[d]	Germany[c]	UK[d]	Poland[d]
1	AS	40.5 (29.1, 50.0)	32.9 (32.4, 33.3)	35.6 (35.0, 36.2)	62.9 (54.0, 70.0)	51.8 (51.5, 52.0)	63.3 (62.7, 63.6)
2	Multi-SNV set	31.8 (21.6, 40.7)	24.7 (23.4, 25.3)	24.4 (22.9, 25.6)	48.4 (36.0, 58.5)	42.1 (38.3, 42.7)	53.1 (51.8, 53.6)
3	Multi-SNV set plus AS	49.3 (39.9, 57.3)	35.0 (31.9, 36.6)	35.3 (31.8, 36.3)	72.2 (65.4, 78.0)	60.8 (24.1, 63.8)	69.0 (58.1, 70.1)

AS, activity score; CYP2D6, cytochrome P450 2D6; SNV, single-nucleotide variant.
[a]Metabolic ratio endoxifen/*N*-desmethyl tamoxifen. [b]Models (variants described in **Table S1**): (1) CYP2D6-AS; (2) SNV sets (including *CYP2D6* SNVs); (3) SNV sets plus CYP2D6-AS. [c]Median R^2 (%) and corresponding 95% confidence interval derived from 20-fold nested cross-validation in the German cohort (500 repeats). [d]Median R^2 (%) and corresponding 95% confidence interval based on predictions derived from model fit in the German cohort with 20-fold (internal) cross-validation (500 repeats).

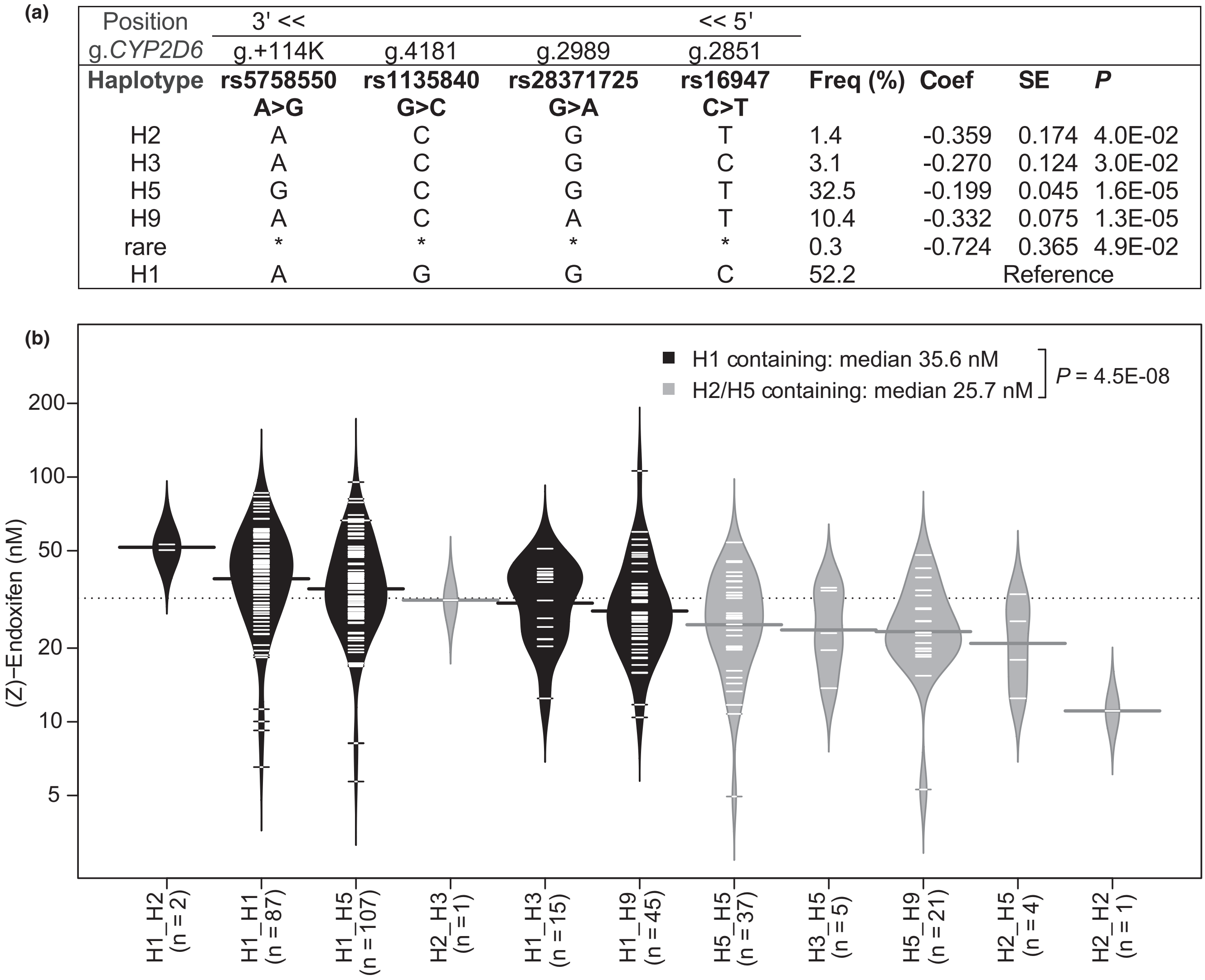
Long-distance haplotypes refine plasma endoxifen prediction in normal metabolizer patients

Since standard *CYP2D6* diplotype assignments cannot satisfactorily explain why low blood endoxifen concentrations occur in patients with normal CYP2D6 metabolizer genotype (AS 1.25 to 2.25), we exemplarily examined the effect of long-distance haplotype assignment on plasma endoxifen prediction. Specifically, we investigated the interdependencies between *CYP2D6**2 variants rs16947 and rs1135840, the *41 defining rs28371725, and their relation to a 114-kb downstream-located variant rs5758550 previously reported to enhance *CYP2D6* promoter activities[33] (**Figure 5**). In a subgroup analysis of combined UM and normal metabolizer patients (CYP2D6-AS of ≥1.25), haplotypes were estimated based on 327 patients of the combined UK and German cohorts. In comparison with the most frequent H1 haplotype (patients with *CYP2D6*1*), all other haplotypes were associated with lower endoxifen concentrations (**Figure 5a**). Diplotypes composed of H5 or H2 haplotypes (presence of rs16947), i.e., patients with *CYP2D6**2 had on average a 29%-reduction of endoxifen concentrations (median 25.7 nM) compared with H1-containing diplotypes (median 35.6 nM; $P = 4.5E{-}08$; **Figure 5b**). Of note, NM/IM patients characterized by haplotypes H9 (*41) or H3 (*10) had either higher or lower median endoxifen concentrations depending on whether they occurred in combination with H1 or H5 haplotypes, respectively. Accordingly, the plasma endoxifen heterogeneity of NM patients can, at least in part, be further resolved by haplotypes composed of >100 kb distantly located variants.

DISCUSSION

This study was motivated by the knowledge gap of the relevant determinants of variable endoxifen concentrations and their pharmacological implications during tamoxifen treatment of patients with hormone receptor-positive early breast cancer. CYP2D6, the key metabolizing enzyme responsible for tamoxifen-to-endoxifen conversion has been suggested as a predictive marker for clinical outcome, yet its usefulness in personalized treatment strategies is controversially debated,[14,36] partly due to its limited predictive power for endoxifen plasma concentration. We performed the first cross-ancestry breast cancer GWAS to investigate additional genetic predictors of plasma endoxifen concentrations that, to the best of our knowledge, represents the largest study in the field. The combined analysis of different ethnic cohorts potentially identifies variants that are more likely to be causal than candidates derived from a purely ancestry-specific approach. We provide strong evidence for multiple associations of loci located within an ~ 1-Mb region at chromosome 22q13 that includes the *CYP2D6* gene. Other recently associated chromosomal regions[22,23] did not replicate in this GWAS, a reason why we consider *CYP2D6* and functionally relevant variants in the surrounding region to be the principal genetic determinants of plasma endoxifen concentrations variability.

The prominence of the 22q13 genomic region is corroborated by an association observed also with metabolic end points MR endoxifen/N-desmethyl tamoxifen, N-desmethyl tamoxifen, (Z)-4-hydroxytamoxifen, and MR (Z)-4-hydroxytamoxifen/



Position	3' <<			<< 5'				
g.*CYP2D6*	g.+114K	g.4181	g.2989	g.2851				
Haplotype	**rs5758550** **A>G**	**rs1135840** **G>C**	**rs28371725** **G>A**	**rs16947** **C>T**	**Freq (%)**	**Coef**	**SE**	***P***
H2	A	C	G	T	1.4	-0.359	0.174	4.0E-02
H3	A	C	G	C	3.1	-0.270	0.124	3.0E-02
H5	G	C	G	T	32.5	-0.199	0.045	1.6E-05
H9	A	C	A	T	10.4	-0.332	0.075	1.3E-05
rare	*	*	*	*	0.3	-0.724	0.365	4.9E-02
H1	A	G	G	C	52.2	Reference		

Figure 5 Long-distance haplotypes refine the subgroup of normal metabolizer patients for stratified endoxifen prediction (UM, NM/NM, NM/IM patients of the combined German and UK cohort; *n* = 327). (**a**) Haplotypes were estimated based on *CYP2D6* variants defining *2 (rs1135840, rs16479) and *41 (rs28371725) as well as downstream enhancer variant rs5758550. Variant positions refer to positions in the *CYP2D6* gene (chromosomal reverse strand) according to The Pharmacogene Variation Consortium (PharmVar). Coefficients (coef) indicating mean differences in log endoxifen between haplotypes compared with the most common haplotype H1 (reference). Haplotype frequencies are given; haplotypes with frequencies <1% were combined into the group "rare." (**b**) Haplotype pairs (diplotypes) were ordered according to their average endoxifen concentrations (nM). The two groups of H1 vs. H5 or H2 containing diplotypes differ significantly in their average endoxifen concentrations (*P* = 4.5E−08). Combinations with "rare" haplotypes, *n* = 2, were excluded from statistical analysis. Of note, H5/H5 diplotypes equivalent to *CYP2D6**2/*2 NM patient status had consistently lower endoxifen concentrations compared with their *CYP2D6**1/*1 counterparts (H1/H1 diplotypes), suggesting incomplete compensation of deleterious allele effects in the former. Plausible candidates for compensation are rs5758550 or the CYP2D6-activity increasing rs1135840_4181G>C as shown for propafenone-5-hydroxylation *in vitro*.[32] Freq, frequency; SE, standard error; NM, normal metabolizer; IM, intermediate metabolizer; UM, ultrarapid metabolizer.

tamoxifen. Genome-wide significant associations outside the *CYP2D6* gene and promoter region strongly support previous findings of expression and metabolite quantitative trait loci in the *CYP2D6* neighboring region.[35,37–39] To study this region in more detail we queried the functional relevance of variants from the extended *CYP2D6* locus. Their partially unlinked genetic relation to *CYP2D6* variants (e.g., $r^2 = 0$ to 0.92 for *CYP2D6*4* across the entire region) suggests a phylogenetic origin prior to the emergence of population-specific *CYP2D6* variants. Whether these *CYP2D6*-flanking variants reflect adaptive mutations[40] has been further investigated via their association with CYP2D6 enzyme activity or expression in an independent human liver bank, and by bioinformatic prediction tools. Two-thirds of the investigated 22q13 candidates were significantly correlated with hepatic CYP2D6 activity or expression. Given the partial absence of linkage with known *CYP2D6* variants we concluded that the *CYP2D6*-flanking region contains regulatory elements that influence *CYP2D6* gene expression. We therefore sought to identify critical regions using *in silico* tools for the prediction of regulatory sites. Our data pinpoint several candidate regions that might harbor this regulatory activity located between ≈170 kb upstream and ≈180 kb downstream of *CYP2D6*, which substantiates previous findings based on genomic/transcriptomic liver studies[35,37] and chromatin conformational capture of long-range interactions between the *CYP2D6* promoter and adjacent regions.[41] We suggest six tentative regions of hotspot clusters comprising potentially functional variants (**Figure 4**) that set the stage for future investigations by functional genomic approaches and by third-generation sequencing to obtain long-range phasing information.[42] Similar to recent findings of a superior prediction of MR endoxifen/N-desmethyl tamoxifen by a deep neural network model based on haplotype phasing from full *CYP2D6* gene sequencing,[24] genomic phasing of the 22q13 region may uncover composite haplotypes affecting CYP2D6 expression via yet unknown enhancer/repressor sites, to potentially improve endoxifen prediction.

To compensate for the current lack of large-scale haplotype data, we applied ensemble machine learning to assess the amount of endoxifen variability explained by our GWAS data, and whether the predictive performance of CYP2D6-AS can be improved by *CYP2D6*-flanking intergenic variants of putative functional relevance. Here, we show that models based on a set of 12 SNVs—including 7 non-*CYP2D6* variants—and combined with CYP2D6-AS enhance the average performance in the prediction of endoxifen or MR endoxifen/N-desmethyl tamoxifen by up to 9% when compared with models considering CYP2D6-AS alone, which reinforces the imperfection of the categorical AS system as previously noted.[3] Notably, known confounders such as ethnogeographic allele frequencies, CYP2D6 inhibitor use, and drug adherence[3] were accounted for; however, the average model performance did not improve considerably when covariates age, weight, or variation in *CYP2C9*, *CYP2C19*, and *CYP3A4*[22] were added. Thus, we conclude that the 22q13 locus including *CYP2D6* is accounting for the bulk of functional genetic variability, contrary to the hypothesis that a clinically relevant (personalized) dosing precision algorithm[22] may be significantly improved by factors independent of *CYP2D6*.

As a proof of concept for long-range genetic interactions in the 22q13 region we showed that the >100-kb distantly linked enhancer variant rs5758550[33] may compensate for the deleterious effect of a CYP2D6-activity reducing effect of gene variant rs16947,[32] as evident from haplotype H5 (**Figure 5**). Given that enhancer variant rs5758550 was not significantly associated when analyzed as a single SNV in the GWAS, the 29% reduction of active metabolite concentrations in NM patients with H2 or H5 diplotypes (median 25.7 nM) as compared with those with H1 diplotypes (median 35.6 nM) strongly supports the notion that functional SNV interactions may exist within the chromosome 22q13 region.

Our study is not without limitations. Contrary to the endoxifen/N-desmethyl tamoxifen end point, the endoxifen prediction models showed no clear improvement in the validation cohorts when the multi-SNV set was added to CYP2D6-AS. Cohort-specific nongenetic factors such as sampling time, seasonality, adherence, and storage conditions not available in our study may have confounded endoxifen prediction. Moreover, rare deleterious variants of ADME genes (Absorption, Distribution, Metabolism, Elimination) known to significantly contribute to genetically encoded functional variability[43] were not captured across the study cohorts; however, their influence accounting for between 6% (CYP2D6)[44] and 11% (across all ADME genes)[43] is far below the missing heritability in our study. As the contribution of other ADME genes at the systemic level is probably minor in comparison with *CYP2D6*,[45] promising candidates for further elucidation of drug variability might thus be factors that locally alter disposition directly in breast tissue.[46] Finally, the 22q13 locus undergoes large-scale rearrangements and microdeletions (causing developmental and neuropsychiatric disorders) which have not been investigated in this study. Comorbidities were exclusion factors in our patients with cancer and, as such, the relevance of these rearrangements in our study is likely minor. However, structural variation as well as epigenetic regulation are both known to affect drug levels and response and should be considered in future studies of this region.

In conclusion, the present study contributes novel data on the improvement of the genome-based prediction of active tamoxifen metabolite levels by showing that the CYP2D6-encompassing chromosome 22q13 region aggregates most if not all of the relevant genetic predictors of variable endoxifen blood concentrations. Specifically, we identified multiple noncoding variants upstream and downstream of CYP2D6 with putative regulatory function. Future pharmacokinetic modeling should incorporate long-range haplotype data of the 22q13 region to extend the prediction of CYP2D6 activity via long-distance genetic interactions. Structural variation, epigenetic modifications, and nonsystemic factors may further elucidate the unexplained portion of variable tamoxifen metabolism.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

FUNDING

This study was supported by the National Medical Research Council Singapore (NMRC/1159/2008, NMRCB1011, NRFCG1516, NMRCG13163, NMRC/CIRG/1423/2015 and MOH-000377) grants and the National Research Foundation of Singapore grants (NRF-NRFI2018-01 and NRF-NRFF2017-02). Additional funding was provided by the Biomedical Research Council, Agency for Science, Technology and Research, Singapore and the Horizon 2020-PHC-2015 grant U-PGx 668353. German funding resources include: Robert Bosch Foundation, Stuttgart, Germany, German Federal Ministry of Education and Research (BMBF 01ZP0502); the German Cancer Consortium (DKTK) and German Cancer Research Center (DKFZ), Partner Site Tübingen, 72074 Tübingen, Germany, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy-EXC 2180-390900677 and EXC 2064/1-390727645 as well as the DFG grants SCHR 1323/2-1 and MU 1727/2-1, and the Interfaculty Center for Pharmacogenomics and Drug Research (ICEPHA) University of Tübingen.

ACKNOWLEDGMENT

Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

M.S. is a member of the European PGx Advisory Board of Agena Bioscience GmbH and has received honoraria for oral presentations at academically organized congresses and meetings, and is editor in *Pharmacogenetics and Genomics* (Editor in Chief), *Drug Research* (Editor in Chief), and *Genome Medicine* (Section Editor). Y.S.Y. has received honoraria from Astra

Zeneca. Y.Z. and V.M.L. are cofounders and shareholders of PersoMedix AB. In addition, V.M.L. is CEO and shareholder of HepaPredict AB. All other authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

W.S., S.W., C.C.K., N.S., B.C., H.B.B., and V.M.L. wrote the manuscript. B.C., C.C.K., S.W., W.S., H.B.B., M.S., and V.M.L. designed the research. M.S., T.E.M., B.G., R.T., K.K., E.S., M.E., W.S., H.B.B., D.E., B.E., W.T., N.K.Z., A.T., B.C., N.S., S.C., J.S.L.L., Z.L., J.L., K.S.S., R.C.H.N., Y.S.Y., E.L., M.W., N.S.W., P.C.S.A., and R.D. performed the research and provided patients. C.C.K., S.W., W.S., R.T., and Y.Z. analyzed the data. S.W., R.T., and Y.Z. contributed new analytical tools.

DATA AVAILABILITY STATEMENT

The genome-wide genotype data that support the findings of this study are available from the corresponding author (B.C.) upon reasonable request. Genome-wide genotype data of the 149 liver samples supporting **Figure S4** and **Data S2** are publicly available from NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/, accession number GSE39036). Genome-wide genotype data from the Polish cohort (supporting **Tables 1, 2**, and **S1**) were downloaded from GEO, accession number GSE129162. Tamoxifen metabolite data of the GWAS cohorts and from the UK derive from previous publication[12] and can be accessed, together with genotype data used in validation, upon request to the corresponding author (W.S.).

1. Early Breast Cancer Trialists' Collaborative Group (EBCTCG) Aromatase inhibitors versus tamoxifen in early breast cancer: patient-level meta-analysis of the randomised trials. *Lancet* **386**, 1341–1352 (2015).
2. Brauch, H. & Schwab, M. Prediction of tamoxifen outcome by genetic variation of CYP2D6 in post-menopausal women with early breast cancer. *Br. J. Clin. Pharmacol.* **77**, 695–703 (2014).
3. Helland, T., Alsomairy, S., Lin, C., Søiland, H., Mellgren, G. & Hertz, D.L. Generating a precision endoxifen prediction algorithm to advance personalized tamoxifen treatment in patients with breast cancer. *J. Pers. Med.* **11**, 201 (2021).
4. Johnson, M.D. *et al*. Pharmacological characterization of 4-hydroxy-N-desmethyl tamoxifen, a novel active metabolite of tamoxifen. *Breast Cancer Res. Treat.* **85**, 151–159 (2004).
5. Mürdter, T.E. *et al*. Activity levels of tamoxifen metabolites at the estrogen receptor and the impact of genetic polymorphisms of phase I and II enzymes on their concentration levels in plasma. *Clin Pharmacol Ther* **89**, 708–717 (2011).
6. Maximov, P.Y. *et al*. Simulation with cells *in vitro* of tamoxifen treatment in premenopausal breast cancer patients with different CYP2D6 genotypes. *Br. J. Pharmacol.* **171**, 5624–5635 (2014).
7. Lash, T.L., Lien, E.A., Sørensen, H.T. & Hamilton-Dutoit, S. Genotype-guided tamoxifen therapy: time to pause for reflection? *Lancet Oncol.* **10**, 825–833 (2009).
8. Desta, Z., Ward, B.A., Soukhova, N.V. & Flockhart, D.A. Comprehensive evaluation of tamoxifen sequential biotransformation by the human cytochrome P450 system *in vitro*: prominent roles for CYP3A and CYP2D6. *J. Pharmacol. Exp. Ther.* **310**, 1062–1075 (2004).
9. Gaedigk, A. *et al*. The CYP2D6 activity score: translating genotype information into a qualitative measure of phenotype. *Clin Pharmacol Ther* **83**, 234–242 (2007).
10. Goetz, M.P. *et al*. Clinical pharmacogenetics implementation consortium (CPIC) guideline for CYP2D6 and tamoxifen therapy. *Clin Pharmacol Ther* **103**, 770–777 (2018).
11. Madlensky, L. *et al*. Tamoxifen metabolite concentrations, CYP2D6 genotype, and breast cancer outcomes. *Clin. Pharmacol. Ther.* **89**, 718–725 (2011).
12. Saladores, P. *et al*. Tamoxifen metabolism predicts drug concentrations and outcome in premenopausal patients with early breast cancer. *Pharmacogenomics J.* **15**, 84–94 (2015).
13. Helland, T. *et al*. Serum concentrations of active tamoxifen metabolites predict long-term survival in adjuvantly treated breast cancer patients. *Breast Cancer Res.* **19**, 125 (2017).
14. Goetz, M.P. & Ingle, J.N. *CYP2D6* genotype and tamoxifen: considerations for proper nonprospective studies. *Clin. Pharmacol. Ther.* **96**, 141–144 (2014).
15. Hertz, D.L. & Rae, J.M. Individualized tamoxifen dose escalation: confirmation of feasibility, question of utility. *Clin. Cancer Res.* **22**, 3121–3123 (2016).
16. Mulder, T.A.M., de With, M., del Re, M., Danesi, R., Mathijssen, R.H.J. & van Schaik, R.H.N. Clinical *CYP2D6* genotyping to personalize adjuvant tamoxifen treatment in ER-positive breast cancer patients: current status of a controversy. *Cancer (Basel)* **13**, 771 (2021).
17. Schroth, W. *et al*. Improved prediction of Endoxifen metabolism by CYP2D6 genotype in breast cancer patients treated with tamoxifen. *Front. Pharmacol.* **8**, 582 (2017).
18. Stearns, V. *et al*. Active tamoxifen metabolite plasma concentrations after coadministration of tamoxifen and the selective serotonin reuptake inhibitor paroxetine. *J National Cancer Inst* **95**, 1758–1764 (2003).
19. Antunes, M.V. *et al*. Influence of CYP2D6 and CYP3A4 phenotypes, drug interactions and vitamin D status on tamoxifen biotransformation. *Ther. Drug Monit.* **1**, 733–744 (2015).
20. Nardin, J.M. *et al*. The influences of adherence to tamoxifen and *CYP2D6* pharmacogenetics on plasma concentrations of the active metabolite (Z)-Endoxifen in breast cancer. *Clin. Transl. Sci.* **13**, 284–292 (2019).
21. Mueller-Schoell, A. *et al*. Obesity alters Endoxifen plasma levels in young breast cancer patients: a pharmacometric simulation approach. *Clin. Pharmacol. Ther.* **108**, 661–670 (2020).
22. Chen, Y. *et al*. Effect of genetic variability in 20 Pharmacogenes on concentrations of tamoxifen and its metabolites. *J Personalized Med* **11**, 507 (2021).
23. Hennig, E.E. *et al*. Non-CYP2D6 variants selected by a GWAS improve the prediction of impaired tamoxifen metabolism in patients with breast cancer. *J Clin Med* **8**, 1087 (2019).
24. van der Lee, M. *et al*. Toward predicting CYP2D6-mediated variable drug response from *CYP2D6* gene sequencing data. *Sci. Transl. Med.* **13**, eabf3637 (2021).
25. Schroth, W. *et al*. Gene expression signatures of BRCAness and tumor inflammation define subgroups of early-stage hormone receptor–positive breast cancer patients. *Clin. Cancer Res.* **26**, 6523–6534 (2020).
26. Copson, E. *et al*. Prospective observational study of breast cancer treatment outcomes for UK women aged 18-40 years at diagnosis: the POSH study. *J. Natl. Cancer Inst.* **105**, 978–988 (2013).
27. Lim, J.S.L. *et al*. Impact of CYP2D6, CYP3A5, CYP2C9 and CYP2C19 polymorphisms on tamoxifen pharmacokinetics in Asian breast cancer patients. *Br. J. Clin. Pharmacol.* **71**, 737–750 (2011).
28. Awada, Z. *et al*. Pharmacogenomics variation in drug metabolizing enzymes and transporters in relation to docetaxel toxicity in Lebanese breast cancer patients: paving the way for OMICs in low and middle income countries. *Omics* **17**, 353–367 (2013).
29. Schröder, A. *et al*. Genomics of ADME gene expression: mapping expression quantitative trait loci relevant for absorption, distribution, metabolism and excretion of drugs in human liver. *Pharmacogenomics J.* **13**, 12–20 (2013).
30. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36**, 1–48 (2010).

31. Polley, E., LeDell, E., Kennedy, C. & Laan, M. SuperLearner: Super Learner Prediction. <https://CRAN.R-project.org/package=SuperLearner> (2019).
32. Zanger, U.M., Momoi, K., Hofmann, U., Schwab, M. & Klein, K. Tri-allelic haplotypes determine and differentiate functionally Normal allele *CYP2D6*2* and impaired allele *CYP2D6*41*. *Clin. Pharmacol. Ther.* **109**, 1256–1264 (2021).
33. Wang, D., Papp, A.C. & Sun, X. Functional characterization of CYP2D6 enhancer polymorphisms. *Hum. Mol. Genet.* **24**, 1556–1562 (2015).
34. Lenk, H.Ç. *et al*. The polymorphic nuclear factor NFIB regulates hepatic CYP2D6 expression and influences risperidone metabolism in psychiatric patients. *Clin. Pharmacol. Ther.* **111**, 1165–1174 (2022).
35. Yang, X. *et al*. Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. *Genome Res.* **20**, 1020–1036 (2010).
36. Hertz, D.L. & Rae, J.M. One step at a time: *CYP2D6* guided tamoxifen treatment awaits convincing evidence of clinical validity. *Pharmacogenomics* **17**, 823–826 (2016).
37. Schadt, E.E. *et al*. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**, e107 (2008).
38. GTEx Consortium The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
39. Schlosser, P. *et al*. Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. *Nat. Genet.* **52**, 167–176 (2020).
40. Sadee, W. The relevance of “missing heritability ” in pharmacogenomics. *Clin. Pharmacol. Ther.* **92**, 428–430 (2012).
41. Wang, D. *et al*. Common CYP2D6 polymorphisms affecting alternative splicing and transcription: long-range haplotypes with two regulatory variants modulate CYP2D6 activity. *Hum. Mol. Genet.* **23**, 268–278 (2013).
42. Kraft, F. & Kurth, I. Long-read sequencing to understand genome biology and cell function. *Int. J. Biochem. Cell Biol.* **126**, 105799 (2020).
43. Ingelman-Sundberg, M., Mkrtchian, S., Zhou, Y. & Lauschke, V.M. Integrating rare genetic variants into pharmacogenetic drug response predictions. *Hum. Genomics* **12**, 26 (2018).
44. Zhou, Y. & Lauschke, V.M. The genetic landscape of major drug metabolizing cytochrome P450 genes—an updated analysis of population-scale sequencing data. *Pharmacogenomics J.* **22**, 284–293 (2022).
45. Puszkiel, A. *et al*. Factors affecting tamoxifen metabolism in patients with breast cancer: preliminary results of the French PHACS study. *Clin. Pharmacol. Ther.* **106**, 585–595 (2019).
46. Gjerde, J. *et al*. Tissue distribution of 4-hydroxy-N-desmethyltamoxifen and tamoxifen-N-oxide. *Breast Cancer Res. Treat.* **134**, 693–700 (2012).