Human genetic determinants of tamoxifen metabolism

Koh, Everett Chier Wei

2020

Koh, E. C. W. (2020). Human genetic determinants of tamoxifen metabolism. Doctoral thesis, Nanyang Technological University, Singapore. <https://hdl.handle.net/10356/150839>

<https://hdl.handle.net/10356/150839>

<https://doi.org/10.32657/10356/150839>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 17:41:46 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**HUMAN GENETIC DETERMINANTS OF
TAMOXIFEN METABOLISM**

EVERETT KOH CHIER WEI

SCHOOL OF BIOLOGICAL SCIENCES

2020

**HUMAN GENETIC DETERMINANTS OF
TAMOXIFEN METABOLISM**

EVERETT KOH CHIER WEI

SCHOOL OF BIOLOGICAL SCIENCES

**A thesis submitted to the Nanyang
Technological University in partial fulfilment
of the requirement for the degree of Doctor of
Philosophy**

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

3 Aug 2020

Date

Everett Koh Chier Wei

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

8 Aug 2020

Date

3 Aug 2020

Date

3 Aug 2020

Date

Prof Ng Huck Hui

Assoc. Prof Koh Cheng Gee

Adj. Prof Khor Chiea Chuen

Authorship Attribution Statement

This thesis does not contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

3 Aug 2020

Date

Everett Koh Chier Wei

Acknowledgements

I would like to express my gratitude to Dr. Khor Chiea Chuen and Prof. Ng Huck Hui for their patient guidance throughout my candidature. I am thankful for the opportunity to work on this project, and also the time spent at the Genome Institute of Singapore. I believe the time and the process has forged me into a better man and has equipped me with the skills to be a better scientist.

I also thank Assoc. Prof. Koh Cheng Gee and my Thesis Advisory Committee members Assoc. Prof. Valerie Lin and Asst. Prof. Melissa Fullwood for the advice and guidance during our meetings. I thank Dr Zheng Li and Sim Kar Seng for sharing their knowledge and expertise when working with the command line. I thank the members of Human Genetics group at GIS for their genotyping expertise and high quality data to analysis. Also, I thank our collaborators at the Singapore National Cancer Centre led by Prof. Balram Chowbay for the tamoxifen metabolite quantification and for sharing their patient collections for this study. I also thank Prof Ng Huck Hui, Dr Winston Chan and Lim Yee Siang of the GIS Stem Cell and Regenerative Biology group for sharing their liver study dataset for the gene expression quantitative trait loci analysis.

To the members of 1-126 Cromwell Road, your companionship all these years has been one of the pillars of my life. To Randy, I am a much better student and thinker because of our conversations.

Finally, to Stephanie and Elisha, finishing this journey would not have been possible without your support. This milestone belongs to the three of us.

Abstract

Germline DNA variants have been found to have an impact on a diverse range of human traits, including the ability to metabolise drugs. In this study, we analysed the association between human genetic variants and serum tamoxifen metabolite levels at steady state in breast cancer patients from Singapore, Germany, and Lebanon. Our study found a strong association between a common genetic marker (rs5751245) on Chromosome 22 with log-transformed serum endoxifen levels ($p = 1.2 \times 10^{-28}$). Because CYP2D6 is a well-known metaboliser of endoxifen, we also conducted genome-wide association for *CYP2D6* expression quantitative trait loci (eQTL). We found a common genetic marker (rs2413670) to be strongly associated with both endoxifen ($p_{endo} = 1.2 \times 10^{-23}$) and CYP2D6 eQTL ($p_{CYP2D6_{eQTL}} = 2.9 \times 10^{-14}$). In combination, rs5751245 and rs2413670 explain up to 30% of the variance observed in steady-state serum endoxifen levels. Our findings further contribute to our understanding of the genetic architecture of endoxifen metabolism, with the potential for translation into better outcomes for estrogen-receptor positive breast cancer patients.

Table of Contents

STATEMENT OF ORIGINALITY	2
SUPERVISOR DECLARATION STATEMENT	3
AUTHORSHIP ATTRIBUTION STATEMENT	4
ACKNOWLEDGEMENTS	5
ABSTRACT	6
TABLE OF CONTENTS	7
LIST OF FIGURES	13
LIST OF TABLES	15
1 INTRODUCTION	17
1.1 Polymorphic metabolism of drugs	17
1.2 Tamoxifen	19
1.3 Tamoxifen and the Estrogen Receptor	20
1.4 Tamoxifen's Metabolism	23
1.5 Tamoxifen's effectiveness is dependent on amount in the bloodstream	27

1.6	Clinical Trial Results of Tamoxifen, <i>CYP2D6</i> genotype, and Breast Cancer	28
1.7	Genome-wide scans identify true positives in complex traits	30
1.7.1	Type 2 Diabetes	31
1.7.2	Amyotrophic Lateral Sclerosis (ALS), a.k.a Lou Gehrig's disease	32
1.7.3	Warfarin Case Study	34
1.8	Relationship between <i>CYP2D6</i> genotype and serum endoxifen concentrations	38
1.9	Implementation of genotyping in clinical testing	42
1.10	Other genetic association tests with tamoxifen and metabolite concentrations and metabolite ratios	44
1.11	Current knowledge gaps	46
1.12	Aims and objectives of this thesis	47
2	METHODOLOGY	49
2.1	Patients and Sample Collection	49
2.2	Measurements of steady state serum levels of tamoxifen and the active metabolite Z-endoxifen	49
2.3	Genotyping Platforms	50
2.4	Software Selection for Processing Genotyping Data and Linear Regression Analysis	51
2.5	Quality Control of Genotyping Data	53
2.5.1	Determining level of missingness in the data	54
2.5.2	Heterozygosity	55
2.5.3	Population Stratification	55

2.5.4	SNP-level QC	58
2.6	Association Testing	60
2.6.1	Linear Regression	60
2.6.2	Visualising the phenotype distribution of our study cohorts	60
2.6.3	Fixed and Random Effects in Meta-Analysis	60
2.6.4	Visualisation of Association Testing	61
2.6.5	Genomic Inflation	61
2.7	Haplotype Phasing	61
2.8	Genotype Imputation	62
2.8.1	Imputation Quality Control	64
2.9	Processing of ChIP-seq raw reads from NCBI Sequence Read Archive (SRA)	64
2.10	Statistical Power Calculation	65
2.11	Interpretation of effect size (β) in the log-transformed linear regression model	65
2.12	Additional Notes	66
3	RESULTS	67
3.1	Patient demographic and genetic ancestry	67
3.2	Data transformation for serum tamoxifen metabolite concentrations	70
3.3	Association testing of <i>CYP2D6</i> Genetic Variants with tamoxifen metabolites	75
3.4	Effectiveness of <i>CYP2D6</i> variants at predicting endoxifen levels	80
3.5	Effectiveness of Activity Score	82

3.6	GWAS of tamoxifen metabolites	87
3.6.1	GWAS of endoxifen	87
3.6.2	GWAS meta-analyses of endoxifen in breast cancer patient cohorts	90
3.6.3	Meta-analysis of GWAS on other tamoxifen metabolites	93
3.7	Fine mapping reveals index SNP rs5751245	96
3.7.1	Effectiveness of rs5751245 at explaining variance in comparison to <i>CYP2D6</i> alleles	98
3.7.2	Conditional analysis on index SNP rs5751245	100
3.8	Endoxifen and <i>N</i>-desmethyltamoxifen SNPs	101
3.9	Conditional analysis on <i>CYP2D6</i>*4, <i>CYP2D6</i>*10, and <i>CYP2D6</i>*41 diplotype	103
3.10	4-hydroxytamoxifen (4OHT) Association Testing	105
3.11	GWAS for eQTL of select genes on Chromosome 22	107
3.11.1	Liver Study (SCRB) Cohort	109
3.11.2	eQTL GWAS results	111
3.12	<i>CYP2D6</i> eQTL effect on endoxifen	120
3.12.1	eQTL analysis conditional on rs5751245 allele dosage	122
3.12.2	Conditional analysis on all possible <i>CYP2D6</i> diplotypes	123
3.13	Endoxifen association testing conditional on rs2413670 allele dosage	125
3.14	SNP location and Open Chromatin Marks in Liver Tissue	127
3.15	Linkage Disequilibrium between endoxifen, <i>CYP2D6</i> eQTL, and <i>CYP2D6</i> allele proxy SNPs	131
3.16	Improved predictive power using endoxifen and <i>CYP2D6</i> eQTL index SNPs	133

3.17	Other <i>CYP2D6</i> alleles in other studies	138
4	DISCUSSION	142
4.1	Genetic variants other than <i>CYP2D6</i> are not associated with steady state endoxifen	142
4.2	GWAS reveals rs5751245 as sentinel SNP for steady state endoxifen	142
4.3	rs5751245 annotated as an <i>TCF20</i> upstream variant	144
4.4	Conclusion	145
4.5	Suggestions for future work	145
5	BIBLIOGRAPHY	147

List of Figures

FIGURE 1: MECHANISM OF ACTION FOR THE SERM, ENDOXIFEN	22
FIGURE 2: METABOLISM OF TAMOXIFEN IN VIVO	24
FIGURE 3: MAIN METABOLIC PATHWAY FOR TAMOXIFEN METABOLISM TO ENDOXYFEN	26
FIGURE 4: GRAPH OF STATISTICAL POWER AND EFFECT SIZE	59
FIGURE 5: IMPUTATION SCENARIO	63
FIGURE 6: GENETIC ANCESTRY OF INDIVIDUAL COHORTS	69
FIGURE 7: PCA OF GENETIC ANCESTRY FOR ALL PATIENTS IN THE TAMOXIFEN STUDY COHORT.	69
FIGURE 8: HISTOGRAM AND KOLMOGOROV-SMIRNOV TEST WITH NORMAL CURVE OVERLAYS FOR ENDOXYFEN	71
FIGURE 9: HISTOGRAM AND KOLMOGOROV-SMIRNOV TEST WITH NORMAL CURVE OVERLAYS FOR 4-HYDROXYTAMOXIFEN	72
FIGURE 10 HISTOGRAM AND KOLMOGOROV-SMIRNOV TEST WITH NORMAL CURVE OVERLAYS FOR N-DESMETHYLTAMOXIFEN	73
FIGURE 11: HISTOGRAM AND KOLMOGOROV-SMIRNOV TEST WITH NORMAL CURVE OVERLAYS FOR TAMOXIFEN	74
FIGURE 12: FOREST PLOTS OF THE THREE MOST COMMON CYP2D6 VARIANTS IN OUR STUDY COHORTS	78
FIGURE 13: LINEAR REGRESSION OF MOST COMMON CYP2D6 VARIANTS IN STUDY COHORTS	81
FIGURE 14: LINEAR REGRESSION OF ACTIVITY SCORE AGAINST $\text{LOG}_{10}(\text{ENDOXYFEN})$	86
FIGURE 15: MANHATTAN PLOT OF INDIVIDUAL COHORT ENDOXYFEN GWAS ..	88
FIGURE 16 QUANTILE-QUANTILE (QQ) PLOT OF P-VALUES FOR INDIVIDUAL BREAST CANCER PATIENT COHORTS	89
FIGURE 17: MANHATTAN PLOT OF ENDOXYFEN META-ANALYSIS	91

FIGURE 18:REGIONAL ASSOCIATION FOR CHROMOSOME 22 LOCUS IN THE META-ANALYSIS OF LOG-TRANSFORMED ENDOXIFEN LEVELS GENETIC MARKER RS5751247 DEMONSTRATES THE STRONGEST EVIDENCE OF STATISTICAL ASSOCIATION IN THE LOCUS. NO GENETIC MARKERS IN CYP2D6 GENOTYPED ON THE GWAS CHIP	92
FIGURE 19: MANHATTAN PLOT OF META-ANALYSES FOR N- DESMETHYLTAMOXIFEN, 4-HYDROXYTAMOXIFEN, AND TAMOXIFEN	94
FIGURE 20: QQ PLOT FOR ALL TAMOXIFEN METABOLITE GWAS META- ANALYSES	95
FIGURE 21 REGIONAL MANHATTAN PLOT OF CHROMOSOME 22:41MB-44MB	97
FIGURE 22: RS5751245 LINEAR REGRESSION PLOTS	99
FIGURE 23: REGIONAL MANHATTAN PLOT OF CHROMOSOME 22:41MB-44MB; CONDITIONAL ANALYSIS ON RS5751245.....	100
FIGURE 24: REGIONAL ASSOCIATION PLOT OF ENDOXIFEN CONDITIONED ON CYP2D6*4, CYP2D6*10, AND CYP2D6*41 DIPLTYPE	104
FIGURE 25: REGIONAL ASSOCIATION PLOT OF CHROMOSOME 22:41MB-44MB FOR 4OHT	106
FIGURE 26: REGIONAL ASSOCIATION PLOT OF RS5751245	108
FIGURE 27 GENETIC ANCESTRY ANALYSIS FOR SCRB COHORT	110
FIGURE 28: GWAS OF 39 GENES EQTL.....	113
FIGURE 29: MANHATTAN PLOTS OF CYP2D6 AND NDUFA6 EQTL GWAS	114
FIGURE 30: QQ PLOTS OF CYP2D6 AND NDUFA6 EQTL GWAS	114
FIGURE 31: BOX PLOT OF GENE EXPRESSION OF GENES IN CHR22:41.6- 43.6MB	115
FIGURE 32: REGIONAL ASSOCIATION PLOTS OF CYP2D6 AND NDUFA6 EQTL GWAS	115
FIGURE 33: REGIONAL ASSOCIATION PLOTS OF CYP2D6 AND NDUFA6 EQTL	116

FIGURE 34: REGIONAL ASSOCIATION PLOTS OF CYP2D6 AND NDUFA6 EQTL CONDITIONAL ANALYSIS	116
FIGURE 35: COMPARISON OF P-VALUES BETWEEN CYP2D6 AND NDUFA6 EQTL GWAS ANALYSIS	117
FIGURE 36: REGIONAL ASSOCIATION PLOTS OF CYP2D6 AND NDUFA6 EQTL CONDITIONAL ANALYSIS	122
FIGURE 37: REGIONAL ASSOCIATION PLOT OF ENDOXIFEN CONDITIONED ON ALL POSSIBLE CYP2D6 DILOTYPES	124
FIGURE 38: REGIONAL ASSOCIATION PLOTS OF CYP2D6 AND NDUFA6 EQTL CONDITIONAL ANALYSIS	126
FIGURE 39: REGIONAL ASSOCIATION PLOTS OF ENDOXIFEN CONDITIONAL ANALYSIS ON RS2413670	126
FIGURE 40: IGV BROWSER VIEW OF RS28371699	129
FIGURE 41: IGV BROWSER VIEW RS5751245	129
FIGURE 42: IGV BROWSER VIEW OF RS2413670 AND RS28371699	130
FIGURE 43: IGV BROWSER VIEW OF RS2413670, RS28371699, AND RS5751245	130
FIGURE 44: PAIRWISE R^2 BETWEEN PROXY SNPS AND CYP2D6 ALLELES	132
FIGURE 45: PAIRWISE R^2 OF SNPS FROM YANG ET AL. (2010)¹⁵⁹ AND CYP2D6*4 AND CYP2D6*10 IN CEU (CAUCASIAN) AND EAS (EAST ASIAN) POPULATIONS FROM LDLINK DATABASE²¹³	140

List of Tables

TABLE 1: COMMON CYP2D6 ALLELES WITH REDUCED ENZYMATIC ACTIVITY	39
TABLE 2: EXAMPLE OF ASSIGNING ACTIVITY SCORE (AS)/METABOLISER STATUS ACCORDING TO CYP2D6 GENOTYPE	39
TABLE 3: POWER CALCULATION	65

TABLE 4: INTERPRETATION OF β WITH RESPECT TO Y	66
TABLE 5: PATIENT COLLECTION DEMOGRAPHICS AND CHARACTERISTICS ..	68
TABLE 6: CYP2D6 ALLELES ASSOCIATION TESTING RESULTS IN OUR STUDY COHORTS.....	79
TABLE 7: ASSOCIATION TESTING RESULTS OF RARE CYP2D6 VARIANTS.....	79
TABLE 8: SUMMARY OF CYP2D6 DILOTYPES AND ACTIVITY SCORE AND COUNTS IN EACH COHORT.....	84
TABLE 9: LINEAR REGRESSION TESTING WITH ACTIVITY SCORE.....	85
TABLE 10: COMPARISON BETWEEN RS5751245 AND MOST COMMON CYP2D6 ALLELE	99
TABLE 11: SNPS WITH NO HETEROGENEITY ($I^2 = 0$) IN BOTH ENDOXIFEN AND NDM ANALYSES.....	102
TABLE 12: CYP2D6*5 ENDOXIFEN ASSOCIATION TESTING RESULTS.....	104
TABLE 13 :TABLE OF SNPS WITH $P < 5E-8$ IN CYP2D6 AND NDUFA6 EQTL	118
TABLE 14: SUMMARY INFORMATION OF INDEX AND PROXY SNPS.....	121
TABLE 15: NATIONAL CENTRE FOR BIOTECHNOLOGY INFORMATION SEQUENCE READ ARCHIVE (NCBI-SRA) DETAILS FOR RAW SEQUENCE FILES USED FROM GSE155908.	128
TABLE 16: LINEAR REGRESSION RESULTS OF RS5751245 AND RS2413670 ALLELE DOSAGE ON ENDOXIFEN AND META-ANALYSIS.....	135
TABLE 17: LINEAR REGRESSION SUMMARY STATISTICS OF RS5751245, THE MOST COMMON CYP2D6 ALLELE IN EACH RESPECTIVE COHORT, AND RS2413670 ON ENDOXIFEN	136
TABLE 18: ALTERNATE ALLELE COUNTS OF RS2413670_C AND RS5751245_C AND FREQUENCY	137
TABLE 19: SNPS FOUND IN OTHER STUDIES	141

1 Introduction

Germline DNA sequence variation impacts a wide variety of human traits such as skin colour, height, weight, as well as susceptibility to diseases. The metabolism of drugs by the human body is also significantly influenced by genetic variation. Drugs undergo a journey of transformation by various enzymes once administered. Reactions are classified into two categories, Phase I (including oxidation, reduction, and hydroxylation), and Phase II (including glucuronidation, sulfation, methylation, and acetylation). In between metabolic steps, drugs also interact with various partners, including transport proteins, receptors, and other small molecules. A large majority of these are gene products, and inherited differences in DNA sequences can affect the pharmacokinetics and pharmacodynamics of drugs.^{1,2}

1.1 Polymorphic metabolism of drugs

In the 1970s, two groups had independently discovered the polymorphic metabolism of drugs. In England, Robert L. Smith and his colleagues found that debrisoquine, an antihypertensive drug, was poorly metabolised by a small but significant group of patients. Furthermore, the poor metaboliser phenotype appeared to be hereditary³. In Germany, Michel Eichelbaum et al. described the same polymorphic metabolism but in the antiarrhythmic agent, sparteine⁴.

A flurry of other reports followed of other drugs that demonstrated the same debrisoquine/sparteine related polymorphic metabolism, such as the tricyclic antidepressant nortriptyline⁵, the cough suppressant dextromethorphan⁶, and antipsychotic perphenazine⁷.

In 1984, Distlerath and Guengerich had isolated the human cytochrome p450 enzyme responsible for the debrisoquine/sparteine metabolism⁸, and we would know this enzyme today as CYP2D6. In 1987, Eichelbaum had mapped the CYP2D6 gene to Chromosome 22q13 of the human genome⁹. Two years later, in 1989, the CYP2D6 gene was cloned and sequenced¹⁰. In 1990, genetic variants that identified the poor metaboliser phenotype were described¹¹. These poor metaboliser genetic variants would later be described as *CYP2D6*4*, *CYP2D6*10*, and *CYP2D6*41*¹².

Using genetics to predict a patient's response to a drug has been a focal point of research and integrated into patient care ever since. Subsequent developments as a result of technological advances that lowered the cost of sequencing and the Human Genome Project contributed to the discovery of more and more genetic variants for *CYP2D6*. In 1996, Daly et al. suggested the adoption of the star allele nomenclature system that is currently still in use today¹².

The first concerted effort to categorise, curate, and maintain the knowledge of all potential genetic variants was in 2001, with the creation of a webpage <http://www.imm.ki.se/CYPalleles/>¹³ (now *defunct*). Today, the Pharmacogene Variation Consortium (PharmVar) (<https://www.pharmvar.org/>) is responsible for curating all the genetic variants of the pharmacologically relevant genes, including CYP2D6¹⁴.

Other resources have emerged such as the Pharmacogenomics Knowledgebase (PharmGKB) (<https://www.pharmgkb.org/>)¹⁵ which curates the information that integrates human genetic variation and its impact on drug

metabolism, and the Clinical Pharmacogenetics Implementation Consortium (CPIC) (<https://cpicpgx.org/>)¹⁶ that publishes guidelines on prescribing and dosing strategies based on genetic makeup. As of 1st June 2020, PharmVar documents 126 distinct *CYP2D6* alleles, excluding subvariants, and the Flockhart Table¹⁷ documents 65 different drugs that are metabolised by CYP2D6.

1.2 Tamoxifen

Tamoxifen was initially named *ICI 46474* and was developed in the 1960s by the Imperial Chemical Industries (ICI) Pharmaceutical Division, now part of AstraZeneca. In 1967, tamoxifen was shown to be an estrogen agonist/antagonist in rat and mouse models¹⁸. Over the next few years, tamoxifen was also found to stimulate ovulation in subfertile women¹⁹, which nearly killed the drug's development, as it was initially developed as part of a contraceptives programme. Tamoxifen's anti-estrogenic properties led to clinical trials as a treatment for advanced breast cancer patients in the early 1970s^{20,21}, where tamoxifen's efficacy and few side effects were noted. Around the same time, it was reported that breast cancers with an abundance of estrogen receptors are more likely to respond to endocrine treatments than breast cancers with less or no estrogen receptors^{22,23}. However, a clever pivot away from contraception to anti-estrogenic applications such as cancer treatment allowed tamoxifen to continue development.

In the mid-1970s, Dr Virgil Craig Jordan and other colleagues demonstrated the anti-tumorigenic properties of tamoxifen on breast cancer^{24–27}. Tamoxifen was already approved for use in metastatic breast cancer in the United

Kingdom and other countries since 1972; only in December 1977 did the United States of America approve tamoxifen's use in metastatic breast cancer²⁸. Clinical trials have consistently demonstrated the advantages of tamoxifen adjuvant therapy in terms of reducing cancer recurrence and cancer mortality^{29–33}.

1.3 Tamoxifen and the Estrogen Receptor

Tamoxifen was subsequently classified as a Selective Estrogen Receptor Modulator (SERM) as it was found to have agonist effects on bone, uterine, and vaginal tissue, but is an antagonist in breast tissue, including breast cancer cells^{27,34}. An agonist is a ligand that, when bound to a receptor, produces a biological response, whereas an antagonist blocks the effects of the agonist. Estrogen stimulates breast epithelial cells to grow and divide. Breast cancer cells proliferate in the same fashion with estrogen stimulation³⁵. Tamoxifen has antagonistic effects in breast tissue, halting proliferation of both normal and cancer breast cells, but is found to have the same stimulating effects of estrogen in bone and uterine tissue³⁶.

There are two types of estrogen receptors (ER), alpha (α) and beta (β). Human tissue expresses both types; however, the amount of each type varies with different tissues. ER α is the predominant form in tissues such as the mammary glands, uterus, bone, and testes, whereas ER β is the prevalent form tissues such as the bladder, ovaries, colon, and the immune system^{37,38}. ER signalling requires the ER monomers to dimerise, and dimerisation involves the binding of a ligand. Different ligands affect the resultant conformation of the monomers, and by extension, the dimer's conformation.

For example, the ER agonist 17 β -estradiol (also referred to as E2) is responsible for normal ER signaling, puts the ER in an “ON” conformation. In contrast, the drug Fulvestrant, known to be a pure ER antagonist, puts the ER in an “OFF” conformation which prevents the monomers from dimerising³⁹. SERMs such as tamoxifen and its metabolites binding to ER monomers result in a conformation that is somewhere in between the “ON” and “OFF” conformation^{40,41}. The differences in conformation alter the ER’s ability to interact with transcriptional co-regulators, as well as affect binding to ER response elements (ERE); both events have an impact on regulating ER-mediated gene expression⁴². ER-signalling pathway is further complicated by the abundances of other ER transcriptional co-regulators, which results in SERMs such as tamoxifen to have both agonist and antagonist properties in different tissues (*Figure 1*).

Constant cycling of ER formation and degradation has to be present in the cell for consistent ER signalling⁴³. The stability of the ER protein is also reliant on post-transcriptional modification such as ubiquitination and is also affected by the ligand that it binds to. Greater stability conferred on the ER-ligand complex discourages ER cycling and hence attenuates ER signalling⁴⁰

Figure 1: Mechanism of action for the SERM, Endoxifen

Estrogen receptors dimerise upon ligand binding. 1) Agonists such as E2 (light blue chevrons) causes a conformational change in the ER and causes it to dimerise, thereby allowing the recruitment of coactivators (yellow hexagons). The complex results in the activation of gene transcription (purple squiggly line) where estrogen response elements (ERE) (pink boxes) in the genome are found. 2) SERMs (green house-shaped pentagon) such as endoxifen similarly causes the dimerization of the ER, and the ability to recruit both coactivators (2a) as well as corepressors (2b). This implies that the mechanism is affected by the relative abundance of the binding partners of ER in different tissues, hence having different responses. 3) Binding of antagonists such as Fulvestrant to the ER results in a conformation that does not allow the ER to dimerise, and hence blocking all downstream effects. Diagram is inspired by Figure 2 of Riggs et al, N Engl J Med, 2003³⁴.

Tamoxifen's main therapeutic metabolites are 4-hydroxytamoxifen (4OHT) and 4-hydroxy-*N*-desmethyltamoxifen (endoxifen), as they have an ER binding affinity of approximately 100-fold greater than tamoxifen; endoxifen and 4OHT have approximately the same binding affinity for ER⁴⁴. Endoxifen and 4OHT have also demonstrated highly similar effects on gene transcription in breast cancer cells in-vitro⁴⁵. However, the concentration of endoxifen in blood serum is on average 6-10 times higher than 4OHT⁴⁶. Therefore, endoxifen is widely accepted as the primary therapeutic compound. Wu et al.⁴⁷ demonstrated in the ER+ MCF7 cell line that endoxifen causes ERα to be targeted for degradation by the proteasome. They also show that endoxifen inhibits ERα-mediated

transcription, thus suppressing estrogen-induced cell proliferation in MCF7 cells. Metabolite concentration also plays a significant role on the ability of tamoxifen to suppress breast cancer cells from proliferating. A high enough endoxifen concentration and endoxifen/estrogen ratio is required for effective suppression of the breast cancer cells⁴⁸.

1.4 Tamoxifen's Metabolism

Tamoxifen is administered orally and is metabolised in the liver into the therapeutic compound endoxifen by a series of cytochrome p450 enzymes (CYPs), Uridine 5'-diphospho-glucuronosyltransferases (UGTs), sulfotransferases (SULTs), and flavin monooxygenases (FMOs) for excretion from the body⁴⁹. *Figure 2* details the transformation of tamoxifen and the enzymes that are known to mediate those reactions.

Tamoxifen's primary metabolism has two routes. The major route is demethylation by CYP3A4/5 into *N*-desmethyltamoxifen (NDM). The minor route is 4-hydroxylation by CYP2D6 into 4OHT⁵⁰. 92% of tamoxifen is metabolised into NDM, whereas 4OHT metabolism only accounts for 7%; the remaining 1% consists of trace metabolites⁴⁶.

The majority of the endoxifen is formed from the secondary metabolism of NDM, which is mediated solely by CYP2D6. A miniscule amount of endoxifen is formed through secondary metabolism of 4OHT⁴⁶. *Figure 3* summarises the transformation of tamoxifen to endoxifen. CYP3A4/5 and CYP2D6 are the main enzymes responsible from the biotransformation of tamoxifen to endoxifen. However, other enzymes such as CYP2B6, CYP2C9, CYP2C19 have also been shown to catalyse these reactions, but at drastically lower efficiencies^{46,51}.

Environmental factors also affect tamoxifen's metabolism into its therapeutic compounds. Stearns et al.⁵² and Jin et al.⁵³ have demonstrated that co-medication of tamoxifen with known inhibitors such as quinidine and paroxetine for CYP2D6 and troleandomycin for CYP3A4 results in drug-drug interactions that lower quantities of the metabolite. Teft et al.⁵⁴ has also demonstrated a seasonal effect related endoxifen levels; increased exposure to sunlight is correlated with increased serum endoxifen levels.

Figure 3: Main metabolic pathway for tamoxifen metabolism to endoxifen

Tamoxifen's main metabolites and enzymes shown. Size and thickness of the arrow lines indicate relative amount of enzyme activity (not drawn to scale). Figure adapted from Borges et al. (2006)⁵⁵

1.5 Tamoxifen's effectiveness is dependent on amount in the bloodstream

Tamoxifen's is prescribed at a standard 20 mg/day. However, the amount of tamoxifen that is metabolised into endoxifen varies widely. Studies that measured circulating endoxifen levels at steady state (i.e. patients have been taking the standard dose for more than >2 months) have observed endoxifen levels less than 5 nM at the lower bounds and greater than 160 nM at the higher bounds^{53,55–57}. Non-compliance on the part of the patient has been cited as a potential reason for the large variance seen when measuring concentrations of metabolites. However, even when non-compliance has been accounted for, large variances in endoxifen levels are still observed⁵⁷.

In-vitro studies by Wu et al.⁴⁷ demonstrated that high concentrations of endoxifen (100 nM) was able to inhibit cell proliferation, even in the presence of estrogen. However, this suppressive effect on cell proliferation decreases at lower (40 and 10 nM) concentrations of endoxifen. Hawse et al.⁴⁸ confirms these concentration-dependent effects of endoxifen and demonstrates that endoxifen exerts its effects by inducing cell cycle arrest, and even apoptosis.

Madlensky et al.⁵⁶ demonstrated female patients with a serum endoxifen level >5.9ng/mL have a 26% lower cancer recurrence rate*. Saladores et al.⁵⁷ demonstrated that patients with lower serum endoxifen concentrations (<14.1 nM) had a higher risk of distant relapse[†] compared to patients with higher serum endoxifen concentrations (>35 nM).

* Also known as disease-free survival. Defined as time from first breast cancer diagnosis to time of second breast cancer event.

[†] Defined as time of diagnosis to time of earliest metastasis or death.

1.6 Clinical Trial Results of Tamoxifen, *CYP2D6* genotype, and Breast Cancer

In 2007, Schroth and Antoniadou et al.⁵⁸ reported association between *CYP2D6* alleles with decreased enzymatic function and worse outcomes in terms of relapse-free time (RFT)[‡] [hazard ratio (HR) = 2.24 (1.16-4.33)[§]; $p = 0.2$] and event-free survival (EFS)^{**} [HR = 1.89 (1.10-3.25); $p = 0.2$]; there was no significant association between *CYP2D6* alleles and overall survival (OS) [HR = 1.73 (0.88-3.41); $p = 0.11$]^{††}. In 2009, Schroth and Goetz et al.⁵⁹ again reported similar findings with RFT [1.57 (1.18-2.08); $p < 0.001$], EFS [1.33 (1.06-1.68); $p = 0.01$], and additionally with disease-free survival (DFS)^{‡‡} [1.29 (1.03-1.61); $p = 0.02$]; no association was again found with OS [1.15 (0.88-1.51); $p = 0.3$]. The two studies mentioned above genotyped the most common alleles of *CYP2D6**4, *CYP2D6**5, *CYP2D6**10, *CYP2D6**41, with Schroth et al. (2009) additionally including *CYP2D6**3 and *CYP2D6**xN alleles.

Other consortia such as the Arimidex, Tamoxifen, Alone or in Combination (ATAC) trial⁶⁰, Breast International Group (BIG) 1-98 trial⁶¹ have demonstrated null association between *CYP2D6* genotype or metaboliser status with breast cancer outcomes ($P > 0.1$). However, these two studies have controversy surrounding their findings. The ATAC trial has been criticized for their study design, as less than 19% of the group of patients randomized to Tamoxifen

[‡] defined as time between surgery to detection of new cancer, either local or distant or contralateral breast cancer

[§] 95% confidence interval indicated in brackets as a range.

^{**} defined as time between surgery and the detection of any new cancer (local, distant, or contralateral), or death from any cause

^{††} Defined as time between surgery and death from any cause.

^{‡‡} Defined as time between first cancer event and second cancer event, or death from any cause.

treatment were genotyped for *CYP2D6* alleles in the end⁶². Furthermore, the authors published in an erratum⁶³ that the genotypes of the *CYP2D6* alleles *CYP2D6*2*, *CYP2D6*4*, *CYP2D6*10*, *CYP2D6*41* were not in Hardy-Weinberg Equilibrium (HWE) amongst the patients that they have genotyped. Genotypes that are not in HWE could introduce potential bias in association tests (more on HWE elaborated in Section 2.5.4). The 1-98 BIG study reported by Megan et al. also has similar criticisms of deviation from HWE amongst genotyped individuals⁶². Additionally, the 1-98 BIG study only has 80% power to study comparatively larger effect sizes ($HR > 1.55$). Another criticism to be levied against Megan et al. is the choice to use tumour DNA for genotyping, where potential aberrations such as deletions or amplifications could be present.

The Austrian Breast and Colorectal Cancer Study Group Trial 8 (ABCSG8) trial⁶⁴ used the measure of disease event (either a recurrence [local or distant], contralateral, second primary cancer, or death from any cause) and in their study, they found that patients carrying two non-functional alleles were significantly at greater risk [odds ratio (OR) = 2.45 (1.05-5.73); $p = 0.04$]. A similar trend was observed for patients with at least one non-functional allele although the association was of not significance [OR = 1.67 (0.95-2.93); $p = 0.07$]. . Although the association between *CYP2D6* genotype and breast cancer outcome is contentious, tamoxifen has consistently proved to be protective against breast cancer relapse and mortality³⁰⁻³³.

1.7 Genome-wide scans identify true positives in complex traits

Mendelian traits or diseases is where the phenotype is usually caused by one gene that has an extremely large effect. Examples of Mendelian traits and diseases includes but is not limited to red-green colour blindness, phenylketonuria, cystic fibrosis, Huntington's disease, and sickle-cell anaemia. Complex traits include spectrum traits such as height, body-mass index (BMI), as well as complex diseases such as Type-2 diabetes, Amyotrophic Lateral Sclerosis (ALS) and hypertension. Drug metabolism is also considered a complex trait as drug pharmacokinetic and pharmacodynamic properties are affected by a multitude of genetic and environmental factors. Complex traits usually having multiple genes and genetic loci contributing a small effect individually, but collectively, is large enough to cause a phenotype². This makes detecting these small effects extremely difficult to distinguish the true associations from the noise when doing association testing.

Before genome-wide scans were affordable, sequencing was a technique that was both expensive and labour intensive. As such, many association studies that were conducted before a genome-wide scan was conducted resulted in many false positives by virtue of cherry-picking candidate genes to study. In the following pages, we explore three examples of how through genome-wide scans, true associations were detected in Type 2 diabetes, Amyotrophic Lateral Sclerosis, and the metabolism of the anticoagulant drug Warfarin.

1.7.1 Type 2 Diabetes

Type 2 diabetes is one such example of a common disease that is influenced by many genetic loci. Each locus contributes a small amount to the eventual manifestation of the disease. Not every individual with the disease will carry all the genetic risk alleles, nor even the risk allele with the greatest effect.

The detection of such small genetic effects requires extremely large sample sizes, only feasible with consortia efforts. As of 2018, 143 risk loci have been discovered in individuals of European ancestry, with a collective number of 62,892 cases and 596,424 controls; the most significant association being rs7903146 ($\beta = 0.3059$, $P = 1.33 \times 10^{-347}$)⁶⁵.

Prior to using genotyping chip arrays that had single nucleotide resolution, genome scans were conducted using microsatellite markers⁶⁶.

The earliest genetic association with Type 2 Diabetes was reported by Meyre et al.⁶⁷ in 2005. The authors reasoned that within the 6q16.3-q24.2 loci (found to be associated to type 2 diabetes in an earlier genome-scan⁶⁸, the gene with the most potential biological effect was the ectonucleotide pyrophosphate phosphodiesterase (*ENPP1*). *ENPP1* plays a role in the insulin signalling pathway⁶⁹. Meyre et al.⁶⁷ studied 1225 cases and 1205 controls in their association study of *ENPP1* variants. However, the association of Type 2 Diabetes with *ENPP1* was never robustly replicated in other larger study populations^{70,71}. The problem faced with the initial study was the inadequate sample size for the small effect size that individual genetic variants exerted on the final phenotype. Later, Grant et al., (2006)⁷² studied a total of 3774 individuals, (1774 cases, 2000 controls) across three populations of European

ancestry and found the association of Chromosome 10q25, which located to the transcription factor gene *TCF7L2*. The association of Chromosome 10q was found to be suggestive in an earlier 2003 study⁷³. This genetic association was subsequently replicated in other study cohorts of across the world [in Iceland, Denmark, United States of America⁷⁴, United Kingdom⁷⁵, Finland⁷⁶, Dutch Breda⁷⁷, Japanese⁷⁸, Mexicans⁷⁹, Asian Indians⁸⁰, Han Chinese⁸¹, Hong Kong Chinese⁸², and Iranians^{83,84}. There is simply no substitute for large sample sizes and rigorous statistical thresholds.

1.7.2 Amyotrophic Lateral Sclerosis (ALS), a.k.a Lou Gehrig's disease

Similar irreproducibility of early findings also affected GWAS studies of neurological diseases such as ALS; the earliest associations published in some of the most prestigious journals for ALS were unable to be replicated. The first GWAS for ALS by Schymick et al.⁸⁵ was published in *Lancet Neurology* in April 2007. However, they did not report any definitive associations of genetic loci with ALS. Subsequently, another study published in the *New England Journal of Medicine* identified *FGGY* as a potential risk loci⁸⁶. A couple of months later, a second but smaller study reported a second potential risk loci with biochemical evidence in *The Lancet Neurology* of the involvement of *ITPR2*⁸⁷. A few months after that in 2008, two reports involved genetic variants near *DPP6* with rather convincing statistics [van Es et al.⁸⁸, Odds Ratio (OR) 1.30 (1.18–1.43), $p = 5.04 \times 10^{-8}$; Cronin et al.⁸⁹, OR 1.37 (1.2–1.56), $p = 2.53 \times 10^{-6}$]. Of note, the patient collections for these studies were rather small, several hundred cases and several hundred controls.

As the years passed, study populations started climbing into the thousands, lending more statistical power, the wheat starts to get separated from the chaff. True associations were starting to float to the top in Chromosome 9q21⁹⁰. Also, previously identified genetic associations were unable to be replicated⁹⁰⁻⁹². Separately, associations in Chromosome 9q21 were replicated in UK Whites⁹³, Finish Whites⁹⁴, and Flanders Belgian Whites⁹⁵.

The underpinning genetic architecture for the association of Chromosome 9q21 with ALS was determined to be an expansion of a hexanucleotide repeat GGGGCC in the non-coding region of *C9orf72*^{96,97}. Association was further confirmed in a Belgian Flanders White population⁹⁸, in a Taiwanese Chinese individuals⁹⁹, in a Brazilian population¹⁰⁰, in a Portuguese population¹⁰¹, and in Ashkenazi and North African Jewish population¹⁰². However, other reports of non-replication were also reported, namely in the Japanese¹⁰³ and in the Korean populations¹⁰⁴, which suggests that *C9orf72* repeat expansion risk in ALS disease pathology differs in the Japanese and Korean populations. Another neurodegenerative disease, Frontotemporal Dementia (FTD), appear to share the same risk loci at *C9orf72*^{96,97}.

It is also important to bear in mind that although *C9orf72* plays a large and important role in ALS disease pathology, its presence is not essential, as only ~40% of familial ALS and ~5% of sporadic ALS patients have the *C9orf72* repeat expansion. Many efforts utilising the genome-wide scan method has both revealed and confirmed the many genes involved in the circuitry of ALS disease pathology, such as *SOD1*, *TARDBP*, *FUS*, and *UBQLN2*. Each loci contributes a small amount to overall disease progression¹⁰⁵.

1.7.3 Warfarin Case Study

One of the most well-documented and actively pursued instances of applying genotype-guided therapy is in the prescription of the oral anticoagulant drug, warfarin. Like many other traits that is presented as a spectrum, blood warfarin level is a complex trait governed by multiple factors. Approved by the United States Food and Drug Agency in 1954, warfarin had been the most commonly prescribed oral anticoagulant drug for the prevention of severe thromboembolic events such as strokes and heart attacks in individuals who are predisposed to hypercoagulative states^{106,107}.

The titration of warfarin dose is a tricky, and highly personal affair. The drug has a rather narrow therapeutic window: too low a dose and the drug would have no beneficial effect on the patient and the risk of thromboembolic stroke remains; too high a dose and the patient would be at risk for (massive) internal bleeding¹⁰⁶. Further complicating the issue at hand is the long half-life of Warfarin. Warfarin is dispensed as a racemic mixture of *R*- and *S*-enantiomers. The *R*-enantiomer has a half-life of 37 to 89 hours and the *S*-enantiomer's half-life is around 21 to 43 hours. For complete clearance of one dose, it takes approximately a week. In cases of mild overdosage, Vitamin K can be administered, however this has consequences on subsequent warfarin treatment, making titration of the drug even more difficult. In cases of severe overdosage, prothrombin complex, clotting Factor VII, or fresh frozen blood plasma may be administered. However, these treatments carry with it the risk of blood-transmitted infections such as

hepatitis and other viral infections, on top of the increase in risk of thromboembolic events¹⁰⁸.

Hence, there was a complicated prescription protocol put in place, and the patient had to be monitored closely for adverse haemorrhagic events. Across the world, doctors use the International Normalised Ratio (INR), which is the ratio between the warfarin dose administered and the time that it takes for the patient's blood to clot¹⁰⁹. Doctors then adjust the patient's dose accordingly to reach and maintain within the therapeutic INR window. This process usually takes approximately a week for the patient to reach a stable INR for measurement¹¹⁰, and takes a median of almost a month to reach the therapeutic INR window¹¹¹. Furthermore, there are contraindications for patients of different ethnicities¹¹⁰.

A Delicate Balancing Act: Vitamin K Dietary Intake and Warfarin Dose

Vitamin K is crucial in the coagulation of blood. It is a required cofactor in the production of prothrombin, clotting Factors VII, XI, and Xa¹¹². Early biochemical studies demonstrated that warfarin acts on the Vitamin K epoxide reductase (VKOR), to produce the anti-coagulation effects¹¹³.

However, warfarin's efficacy is also affected by the amount of dietary Vitamin K that a patient consumes. Vitamin K is commonly found in green, leafy vegetables, as well as purple-pigmented fruits¹¹⁴. Diet plays a role in the amount of warfarin that is required to maintain a stable INR^{115,116}. The United States National Institute of Health Drug-Nutrient Task Force recommends a stable intake of Vitamin K-rich foods if it is part of a patient's regular diet¹¹⁴. Recent

studies have demonstrated that warfarin is in fact a competitive inhibitor of the Vitamin K binding site on the VKOR protein¹¹⁷, going a long way to explain why diet has such a great effect on warfarin efficacy.

Pharmacokinetics of Warfarin in Humans

As mentioned in the paragraphs above, warfarin is dispensed as a racemic mixture containing both *R*- and *S*-enantiomers of warfarin. Biochemical studies in both rats and humans demonstrated *S*-warfarin to be the more potent of the two enantiomers^{118–120}. Metabolism of warfarin deactivates its vitamin K inhibition properties^{121,122}. In humans, warfarin is metabolised by a handful of Cytochrome P450 enzymes, including CYP1A1, CYP1A2, CYP2C9, CYP2C19, and CYP3A4. Of the lot, only CYP2C9 shows to have activity on *S*-warfarin^{122,123}.

Genetic Association Studies on Warfarin Dose

Early genetic associations studies prior to the widespread availability of genome-wide chip arrays were predominantly candidate gene studies. Extrapolating from known biochemical interaction information, researchers first established the genetic association between warfarin and its metabolising enzyme CYP2C9¹²⁴. Later, it was also shown that patients carrying loss-of-function variants of *CYP2C9* (codified by *CYP2C9**2 [rs1799853; C430T, Cysteine-144-Arginine amino acid substitution] and *CYP2C9**3 [rs1057910; A1075C; Isoleucine-359-Leucine amino acid substitution]) required extremely low doses of warfarin to reach the therapeutic INR¹²⁵. For these patients, the administration of regular doses of warfarin may prove to be fatal.

A pair of studies published in 2004 established the link between *VKOR* to warfarin-resistance in rats and also found the location of the gene in humans^{126,127}. Subsequently, an unbiased, genome-scan approach was applied and found strong association of *VKORC1* (Vitamin K epoxide reductase complex subunit 1) genetic variants with warfarin dosing in humans¹²⁸. The association of *VKORC1* and *CYP2C9* was subsequently confirmed using genome-wide chip array methods that scanned the entire genome in an unbiased fashion^{129,130}.

The sentinel SNPs for *VKORC1* are rs9923231 and rs9934438. rs9923231 is an upstream, promoter region SNP for *VKORC1*, and rs9934438 lies on the first intron of *VKORC1*, and they are in almost perfect linkage-disequilibrium with each other¹³¹. The associations of these two genes was widely replicated in many populations across the globe (summarised in the meta-analysis and review by Jorgensen et al.¹³¹).

The identification of *VKORC1* and *CYP2C9* genotypes influencing warfarin dosing led to the establishment of genotype-guided dosing regimens that more accurately predicted the patients' required dose¹³². However, scepticism of genotype-guided dosing regimens remains due to the less than modest effects seen in one study¹¹¹ and in another, no statistically significant difference between the genotype-guided strategy and the conventional clinically-guided strategy¹³³.

Warfarin: A Conclusion

The success of a genotype-guided strategy is greatly limited in the case of warfarin dosing, because in the real world, doctors are not able to control what their patients eat on a day-to-day basis. The studies that measured the the

effectiveness of a conventional dosing strategy versus a genotype-guided dosing strategy does not account for the dietary covariate, and hence found little to no difference in outcome^{111,133}. Nonetheless, the evidence that warfarin dosing is clearly effected by genotype is beyond dispute¹³¹, and provides a solid foundation for the recommendation and progress plan warfarin dosage adjustment.

1.8 Relationship between *CYP2D6* genotype and serum endoxifen concentrations

Currently, there are more than a hundred different variants of *CYP2D6* alleles catalogued^{15,134–136}. The most common variants [defined as alleles with a minor allele frequency (MAF) < 5% in any ethnic cohort], their proxy SNP, and the mutation that characterises the allele are listed in [Table 1](#). Gaedigk et al.¹³⁷ suggested in 2008 that genotype association testing with endoxifen levels adopt the activity score method; [Table 2](#) demonstrates how each individual's activity score and metaboliser status is determined by genotyping. However, legacy terminology such as metaboliser status (Poor, Intermediate, Normal, Ultra) is still in use in 2020, as demonstrated in PharmGKB's documentation¹³⁸.

Table 1: Common *CYP2D6* alleles with reduced enzymatic activity

<i>CYP2D6</i> allele	Defining Name/Change	dbSNP_ID	Comments	Enzyme Activity	Allele Classification
<i>CYP2D6</i> *1			Wild-type	normal	Extensive Metaboliser (EM)
<i>CYP2D6</i> *4	1846C>T	rs3892097	splicing defect	none	Poor Metaboliser (PM)
<i>CYP2D6</i> *5	whole-gene deletion	NA	<i>CYP2D6</i> gene deleted	none	PM
<i>CYP2D6</i> *6	1707delA	rs5030655	118 frameshift	none	PM
<i>CYP2D6</i> *9	2615-2617delTTC	rs5030656	p.K281del	decreased	Intermediate Metaboliser (IM)
<i>CYP2D6</i> *10	100G>A	rs1065852	p.P34S	decreased	IM
<i>CYP2D6</i> *15	137insA; 137_138insA		in-del; 46 frameshift	none	IM
<i>CYP2D6</i> *17	1023G>A	rs28371706	p.T107I	decreased	IM
<i>CYP2D6</i> *29	1659G>A	rs61736512	p.V136M	decreased	IM
<i>CYP2D6</i> *41	2988C>T	rs28371725	aberrant splicing	Decreased	IM
Table extracted from https://www.snpedia.com/index.php/CYP2D6 , accessed on 27 Dec 2019					

Table 2: Example of assigning Activity Score (AS)/Metaboliser Status according to *CYP2D6* genotype

Individual	Allele 1	Allele 2	Genotype	Genotype Category	Allele 1 AS	Allele 2 AS	Total AS	Metaboliser status
1	*1	*1	*1/*1	EM/EM	1	1	2	Normal
2	*1	*10	*1/*10	EM/IM	1	0.25	1.25	Normal
3	*4	*4	*4/*4	PM/PM	0	0	0	Poor
4	*4	*41	*4/*41	PM/IM	0	0.5	0.5	Intermediate
5	*1xN	*10	*1xN/*10	UM/IM	2	0.25	2.25	Ultra

The method to determine each *CYP2D6* allele is by genotyping the *CYP2D6* allele's proxy SNP. For example, for *CYP2D6*10* is rs1065852, for *CYP2D6*4* is rs3892097, for *CYP2D6*41* is rs28371725. A more comprehensive and correspondingly more expensive method is genotyping using a microarray such as the Roche AmpliChip CYP450. An exhaustive list of alleles may be found on PharmVar's website <https://www.pharmvar.org/gene/CYP2D6>¹³⁶.

Association studies between the *CYP2D6* genotype and tamoxifen metabolites phenotypes were first reported in 2011. Mürdter et al.⁵¹ genotyped poor metaboliser (PM) alleles (*CYP2D6*3*, *CYP2D6*4*, *CYP2D6*5*, *CYP2D6*6*, *CYP2D6*7*, *CYP2D6*8*) and intermediate metaboliser (IM) alleles (*CYP2D6*9*, *CYP2D6*10*, *CYP2D6*41*) and used the genotype information to assign phenotype, (PM/PM, PM/IM, PM/EM, EM/EM, et cetera.) Allele assigning was performed according to guidelines laid out by see Zanger et al.¹³⁹). They performed a linear regression analysis on square root transformed endoxifen concentrations against metaboliser phenotype. However, the implementation of the linear regression is also flawed as they present metaboliser status as an ordinal variable. As an example, IM/IM and PM/EM have cumulative AS of 1, but the interpretation of *Figure 4b* in their paper suggests that PM/EM is higher ranked than IM/IM, which will inevitably skew the regression line. Furthermore, the authors did not report an effect size (β) for the linear regression performed. Hence, the reported variance explained ($R^2 = 0.38$ for levels of endoxifen) should be taken with a pinch of salt as it is not directly comparable with other measures of variance explained from a regression.

Madlensky et al.⁵⁶ found that age, ethnicity, Body-Mass Index (BMI) and square root-transformed serum tamoxifen concentrations explained approximately 28% of the variance observed in serum endoxifen levels. The addition of the *CYP2D6* phenotype (metaboliser status i.e. UM, EM, IM, or PM [see [Table 2](#)]) explained a further 18% of the variance. It is unclear as to how the metaboliser status was implemented in the analysis of variance, as such, it is difficult to compare the exactly the variance explained by *CYP2D6* alleles.

Teft et al.⁵⁴ found that *CYP2D6* genotypes (e.g. PM/PM, EM/EM...) explained ~30% of the variance in natural logarithm transformed-endoxifen levels in their study, although they only genotyped the common alleles (PM alleles: *CYP2D6**3, *CYP2D6**4, *CYP2D6**5; IM alleles: *CYP2D6**9, *CYP2D6**10, *CYP2D6**41). There are a couple of criticisms on their model: firstly, the reported 30% also includes age and BMI, and a fixed intercept value for the linear regression; secondly, it is unclear as to how the categorical variable of the *CYP2D6* genotype EM/EM or IM/PM has been implemented in this linear regression. Also, *CYP2D6* PM/PM was not included in their model, according to [Table 3](#) of the paper.

Saladores et al.⁵⁷ found that *CYP2D6* allele-derived Activity Scores (AS) explained between 33%-38% of the variance observed in square root-transformed serum endoxifen concentrations. However, the AS score assignment is dissimilar to current standard. The authors genotyped PM alleles *CYP2D6**3, *CYP2D6**4, *CYP2D6**5, and *CYP2D6**6, IM alleles *CYP2D6**9, *CYP2D6**10, *CYP2D6**41, and duplicate alleles. From the above genotypes, they classified AS using the

following scheme: 0 (PM/PM), 0.5 (PM/IM), 0.75 (IM/IM), 1 (PM/IM), 1.5 (IM/EM), 2 (EM/EM), 3 (UM/EM).

Schroth et al.¹⁴⁰ reported the overwhelming significant association between *CYP2D6* genotype derived activity score and endoxifen concentration ($10^{-4} < p < 10^{-21}$) as well as association with endoxifen/NDM ratio ($p < 10^{-15}$). They surveyed 11 *CYP2D6* alleles (5 non-functioning [**3*, **4*, **5*, **6*, **7*], 3 partially functioning [**9*, **10*, **41*], and 3 alleles for duplication [**1xN*, **2xN*, **35xN*]). They found that *CYP2D6* genotype-derived activity score explained a median 39-58% of the variance observed in endoxifen concentrations in the regression modelling of their study cohorts. This study downgraded the AS for *CYP2D6*10* to 0.25 from the previous guidelines of 0.5 stipulated in 2008¹³⁷.

In summary, the above studies that have attempted to demonstrate *CYP2D6* genotype and correlation to the phenotype of serum endoxifen levels are not comparable to each other. Either due to differences in data transformation, or categorisation of *CYP2D6* genotype, or from ambiguity of how the regression was performed. It is also worthy to note that neither of the five abovementioned studies reported an effect size (β) for the regression models presented.

1.9 Implementation of genotyping in clinical testing

The approaches typified by Schroth et al.^{58,59,140} is the standard practice in clinical testing¹⁴¹ with varying degrees of comprehensiveness depending on which centre is surveyed. With this genotyping framework, the CPIC guidelines for the classification of *CYP2D6* genotype activity score and recommendations for dosing/alternative therapies was developed^{16,142}.

Firstly, not all alleles of the same classification are phenotypically equal. For example, the *CYP2D6**5 allele is a deletion of the gene from one of the chromosomes. For this reason, *CYP2D6**5 is classified as a non-functional allele, and rightfully given an activity score of zero. However, *CYP2D6**4 is also classified as a non-functional allele. *CYP2D6**4 (rs3892097 1846C>T) results in the creation of an alternate splice site in exon 4 of the mature *CYP2D6* mRNA (splice acceptor site from position 1846 shifts to position 1882). Sangar et al.¹⁴³ demonstrated that the resultant protein product is not a truncated polypeptide, but a shorter protein product that is missing 12 amino acids, from position 169 to 180. The protein product of rs3892097 does indeed demonstrate attenuated enzymatic activity but does not deserve an activity score of zero.

Secondly, inter-ethnic differences in allele frequency is well noted, such as the high frequency of *CYP2D6**10 in the East Asian populations, and the high frequency of *CYP2D6**17 and *CYP2D6**29 in the African populations^{134,144}. The issue of allele frequencies leads to the question of costs when it comes to pharmacogenetic testing; either survey a large panel of *CYP2D6* alleles, which will include the most common alleles for all the population cohorts or genotype a handful and potentially miss important alleles for the individual. In this case, we cannot have the cake, and eat it too. An example for *CYP2D6* pharmacogenetic testing for *CYP2D6* costs US\$479, testing for 25 variants¹⁴⁵.

Thirdly, rare variants and *de novo* variants would be missed out using this testing methodology. In the study of other complex diseases, rare variants can sometimes have large effects on the resultant phenotype. Oftentimes, these studies employ next generation sequencing and large sample sizes of a few

thousand individuals, both for the occurrences of these rare variants to be frequent enough for statistical relevance, as well as to have the adequate statistical power to declare meaningful associations. None of the studies investigating tamoxifen and its metabolite levels have been larger than a thousand. Only one study by Sutiman et al.¹⁴⁶ investigating the downstream metabolism of 4OHT and endoxifen by UGTs has employed exon sequencing in 240 individuals.

1.10 Other genetic association tests with tamoxifen and metabolite concentrations and metabolite ratios

CYP2B6 metabolises tamoxifen to 4OHT.⁴⁶ *CYP2B6*6*, an allele of decreased enzymatic function¹⁴⁷, was tested for association with tamoxifen and the primary metabolites 4OHT, NDM and endoxifen^{51,54} but no association was found between metabolite concentration and genotype.

CYP2C9 is able to metabolise tamoxifen to NDM, as well as to 4OHT, although to a significantly lesser extent as compared to CYP3A4/5 and CYP2D6 respectively¹⁴⁸. *CYP2C9*2* and *CYP2C9*3* alleles are non-functional alleles¹⁴⁹. Association testing for *CYP2C9*2* and *CYP2C9*3* have led to contrasting results: Teft et al.⁵⁴ and Jin et al.⁵³ found no association between *CYP2C9* alleles and concentrations of tamoxifen or its metabolites. Lim et al.¹⁵⁰ also found no association between *CYP2C9*3* and tamoxifen nor its metabolites concentrations. However, Mürdter et al.⁵¹ and Saladores et al.⁵⁷ found significant association ($p < 0.005$) between *CYP2C9*2* and/or *CYP2C9*3* with lower 4OHT and endoxifen concentrations when compared to wild-type (*CYP2C9*1/*1*).

CYP2C19 metabolises tamoxifen to 4OHT as well. *CYP2C19*2* and *CYP2C19*3* are non-functional variants¹⁴⁹. Lim et al.¹⁵⁰ and Mürdter et al.⁵¹ both found no association between the two non-functional *CYP2C19* alleles and tamoxifen or its metabolite concentrations, or the metabolite concentration ratios.

CYP3A4 and CYP3A5 enzymes metabolise tamoxifen to NDM and 4OHT to endoxifen. *CYP3A4*22* has decreased function, whereas *CYP3A5*3* and *CYP3A5*6* have no metabolic function. Teft et al.⁵⁴ found *CYP3A4*22* to be associated with higher concentrations of endoxifen ($p < 0.05$) with concomitant higher concentrations of tamoxifen ($p < 0.001$). No association was found for *CYP3A5* alleles^{46,53,151}.

Genome-wide association studies have been carried out for tamoxifen, but not in the context of steady state metabolite levels. Kiyotani et al.¹⁵² performed a genome-wide scan and described a loci on Chr10q22 to be associated ($p = 2.02 \times 10^{-4}$) with recurrence free survival in Japanese patients treated with tamoxifen. However, this p-value threshold does not surpass the accepted genome-wide significance threshold^{153–155} of $p < 5 \times 10^{-8}$. Onishi et al.¹⁵⁶ performed a genome-wide scan in patients undergoing tamoxifen monotherapy for 14-28 days, prior to surgery. The authors tested for Ki-67 response in the excised breast cancer tumor after tamoxifen therapy, and found three loci, rs17198973 on chromosome 4q34.3, rs4577773 on Chr6q12, and rs7087428 on Chr10p13, all of which did not surpass genome-wide significance ($p = 5.6 \times 10^{-6}$).

Hennig et al. (2019)¹⁵⁷ performed an genome-wide association test against the metabolic ratio calculated from the concentration of endoxifen

divided by the sum of the 13 other tamoxifen metabolites [described in Hennig et al (2015)¹⁵⁸]. They genotyped 192 patients out of the 342 patients they had recruited; 96 of the top ranking MR ratios, and 96 of the bottom ranking MR ratios. The authors set the cutoff MR previously determined in Hennig et al. (2015)¹⁵⁸ and employed the Cochran-Armitage trend test for association testing. They found 7 SNPs to be significantly associated with threshold MR, that surpassed the genome-wide significance threshold, all of which on Chromosome 22 and within a 500 kb window, flanking the *CYP2D6* loci. They did not perform association testing with just endoxifen levels. Although the authors claim increased performance of the predictive potential using MR (their models claim to be able to explain 50% of the variance in MR)¹⁵⁸, it is not practical in real life to determine tamoxifen metabolite concentrations of 14 different metabolites. Also, the association testing they conducted was based on an arbitrarily derived threshold MR value, although based on the cutoff 6 ng/mL threshold suggested by Madlensky et al.⁵⁶, the correlation between endoxifen and MR was not even near perfect (R^2 between endoxifen and MR not reported).

1.11 Current knowledge gaps

To date, no hypothesis-free genome-wide association test has been conducted for steady state serum tamoxifen concentrations as a quantitative trait. Neither has it been done for the concentrations of the primary metabolites NDM, 4OHT, or endoxifen. Although *CYP2D6* alleles have been demonstrated to be strongly associated with varying metabolite concentrations, other genetic factors could also potentially affect the concentrations of tamoxifen and its metabolites in the blood serum at steady state.

Also, no genetic variant that is common and applicable across all ethnic populations has ever been identified to be associated with tamoxifen, or its metabolites. A hypothesis-free genome wide scan is able to identify genetic variants that are common across ethnic cohorts that are strongly associated with tamoxifen and its metabolites' concentrations.

Although *CYP2D6* gene expression has been linked to several genetic variants and these same variants have also been shown to demonstrate association with dextromethorphan metabolism¹⁵⁹, these associations have not been demonstrated in tamoxifen. *CYP2D6**2 and *CYP2D6**17 proxy SNPs have also been demonstrated to have an impact on *CYP2D6* expression in HepG2 cells¹⁶⁰ but not in human liver tissue.

1.12 Aims and objectives of this thesis

I am adopting a hypothesis-free approach by conducting a genome-wide scan for genetic variants that are associated with steady state tamoxifen metabolite levels. Alongside, we will be comparing our results to variants that are known to be associated with tamoxifen metabolism, namely the *CYP2D6* alleles.

I aim to identify genetic variants that are common across the multi-ethnic cohorts that are associated with tamoxifen metabolites' concentrations. Furthermore, I aim to explain the contribution of *CYP2D6* gene expression in the liver tissue alongside other genetic contributions to variations in serum tamoxifen metabolites' concentrations. I also aim to map the chromatin landscape in the liver tissue of the loci that are associated with these traits.

The objective is to identify genetic variants that explain serum tamoxifen metabolites' concentrations as well as *CYP2D6* gene expression and the reconcile these genetic variants that are associated with both traits. Next, by placing these genetic variants in the context of chromatin accessibility in the physiologically relevant tissue, I intend to make conclusions on the genetic architecture for tamoxifen metabolism. Following which, I will compare the genetic variants found from my hypothesis-free genome-wide scan to the classical *CYP2D6* allele-derived Activity Score method of predicting endoxifen concentrations to determine its effectiveness.

2 Methodology

2.1 Patients and Sample Collection

A total of 512 patients from three different ethnic groups were included in the tamoxifen metabolite study. From Singapore, 152 female patients who were histologically diagnosed with ER-positive breast cancer and of Chinese ethnicity were included from the cohort previously described in Lim et al.¹⁵⁰. Another 70 pre-menopausal female patients with ER-positive breast cancer were recruited from the American University of Beirut Medical Center¹⁶¹. 290 female post-menopausal ER-positive breast cancer patients were selected from the tamoxifen arm of a prospective, observational multicentre trial (IKP211; German Registry of Clinical Studies DRKS 00000605)⁵¹.

Institutional ethics approvals were obtained from the ethics committee of the National Cancer Centre Singapore (Ref:2008/402/B) (Singapore), the AUB institutional ethics review board (Lebanon), the Medical Faculty of the University of Tübingen (Ref:340/2004).

All patients were given an oral dose of 20 mg of tamoxifen daily.

2.2 Measurements of steady state serum levels of tamoxifen and the active metabolite Z-endoxifen

Patients' blood samples (at least 3 ml of blood) were collected at least eight weeks post commencement of treatment. Plasma extraction was performed by centrifugation under light protection within 30 minutes of venepuncture and stored at -80°C until analysis. Steady-state plasma concentrations of tamoxifen and its primary metabolites, *N*-desmethyltamoxifen (NDM), 4-hydroxy-

tamoxifen (4OHT) and the therapeutic compound 4-hydroxy-*N*-desmethyltamoxifen (endoxifen)(ENDO) in their active (*Z*)-isomers, were quantified using liquid chromatography tandem mass spectrometry in the multiple reaction monitoring mode performed on an Agilent 1200 Series Rapid Resolution LC System coupled to a 6460 triple quadrupole mass spectrometer (Agilent Technologies, Waldbronn, Germany). Protocol for measuring tamoxifen and its primary metabolites were previously described by Mürdter et al.⁵¹. In brief, 50 µl of plasma was added to 100 µl of 1% acetic acid in acetonitrile containing the internal standards for protein precipitation. Following centrifugation, the supernatant was added to 150 µl of 0.1% acetic acid in water. Chromatographic separation of analytes was achieved using a ZORBAX Eclipse plus C18 (Agilent Technologies, Waldbronn, Germany) with gradient elution using acetonitrile in 0.1% acetic acid in water at a column temperature of 45°C. Concentrations below the lower limit of quantification (LLOQ) were taken to be half the LLOQ.

2.3 Genotyping Platforms

Genome-wide genotyping was carried out using Illumina's Infinium OmniExpress 12 BeadChip platform, and the replication Singaporean cohort was genotyped on the Infinium OmniExpress 24 Beadchip platform, according to the manufacturer's instructions. These Beadchips are able to analyse more than 700,000 common single nucleotide polymorphisms (SNPs) representative of the entire human genome^{162,163}.

To increase the resolution of the genome-wide scan, we carried out whole genome sequencing (WGS) using Illumina's Next Generation Sequencing

technology¹⁶⁴ to an average depth of 30X, with 80% of target bases covered 20X or more. The WGS genotypes can also be used as a reference panel for imputation into datasets that were genotyped using the BeadChip arrays. The process of imputation is further elaborated in [Section 2.8](#).

Single SNP genotyping for the *CYP2D6* alleles was performed using the TaqMan Real-Time PCR assay¹⁶⁵.

2.4 Software Selection for Processing Genotyping Data and Linear Regression Analysis

As serum endoxifen is a continuous variable, the appropriate statistical test for association testing is the linear regression, assuming that the effect of allele dosage is additive (i.e. with every increasing copy of the minor allele, there is a concomitant increase or decrease of the phenotype analysed).

Software packages that are able to conduct a linear regression analysis such as the IBM SPSS, or with R¹⁶⁶ can also be used for genome-wide association studies (GWAS). However, commercial or stock solutions respectively are not optimised to analyse several hundred thousand variants across study cohorts in the thousands. Certain research groups have developed custom R packages such as GWASTools¹⁶⁷ and snpStats¹⁶⁸ to streamline the process of analysing GWAS datasets. Custom C++ software for the sole purpose of handling genome-wide association data such as PLINK^{169,170}, and SNPTEST¹⁷¹ have also been developed.

PLINK and SNPTEST both operate in the command line environment, which enables the employment of high-performance computing server computers,

that are faster and more efficient than their desktop counterparts. The server environment enables parallelisation of workflows to be implemented through multiple job submissions. PLINK has been the preferred software in this study for its inclusion of "data management, summary statistics, population stratification, association analysis, and identity-by-descent estimation"¹⁶⁹ all in one software package. Furthermore, PLINK in its latest iteration (version 1.90)¹⁷⁰ enables the use of multicore, multithreaded operations which further reduces the time taken for the most computationally demanding tasks, such as identity-by-descent calculations and multivariate regression analysis. The SNPTEST suite of tools developed by Marchini et al¹⁷¹ segregates the functions independently, requiring the whole suite to be used at different times: QCTOOLS¹⁷² has to be used for population stratification, summary statistics, and identity-by-descent calculations; SNPTEST¹⁷¹ for the association analysis; GTOOL¹⁷³ for the data management. Furthermore, there is no discernible difference between the association results output by SNPTEST and PLINK¹⁷⁴.

Another distinct advantage of custom solutions such as PLINK and SNPTEST is the utilisation of binary file formats for encoding genotype data (*.bed* files for PLINK, *.bgen* files for SNPTEST). Binary file formats are significantly more efficient in disk-space usage, as well as in read-write performance for any operation in the workflow¹⁷⁵. The disadvantage of this particular file format approach is the segregation of sample, variant, and genotype information into separate files. Other text-based formats such as the *Variant Call Format* (VCF) combines sample, variant, and genotype data into a single file. Although the VCF format is convenient for the purposes of file-sharing between

collaborators and consortiums, it is not the ideal file format when carrying out the data processing.

We used IMPUTE2 (version 2.3.1)¹⁷⁶ for fine-mapping the genomic area of interest by imputation. Pre-imputation haplotype-phasing from known genotypes was carried out using SHAPEIT (version 2.12)¹⁷⁷. The other major imputation software used by the genetics community is BEAGLE^{178,179}. IMPUTE2 and BEAGLE have historically demonstrated similar imputation accuracies, with IMPUTE2 even edging out the competition when imputing common variants ($MAF > 5\%$)¹⁸⁰, although BEAGLE has demonstrated higher throughput efficiencies¹⁸¹. We eschewed using BEAGLE because it is built on Java and we had previously faced difficulties with incompatible Java libraries when utilising the high-performance computing cluster at the Genome Institute of Singapore.

2.5 Quality Control of Genotyping Data

Only autosomal variants were analysed in our study, assuming that metabolism of a drug is not influenced by sex-linked genetic variants. This assumption was made with the full awareness that breast cancer is a disease that predominantly affects women (approximately 99% Female, 1% male)^{182,183} and our patient collection includes only female patients.

Genotype-calling inaccuracies due to poor input DNA quality leads to a series of problems, including marker exclusion due to genotype calling errors, spurious population stratification determination, and false associations. Hence, only individuals with high quality DNA samples genotyped to a high completion rate were retained for analyses.

Genome wide association studies typically includes minimally hundreds of thousands, to millions of SNPs (84.7 million SNPs in the 1000 Genomes Project¹⁸⁴), each SNP having the potential to be associative with the phenotypic trait in question. Because of the sheer number of potential markers involved, prudence is necessarily exercised such that true associations do not fall through the cracks due to poor genotyping practice, but also ensure that we keep the number of spurious associations that may be followed-up on to a minimum and not squander finite resources chasing artefacts¹⁸⁵.

Samples were first stratified per collection for the Singaporean (SGP), Lebanese (LBN) and German (DEU) cohorts prior to carrying out any quality control steps and regression analyses. This is done to account for the variance in minor allele frequencies observed in different populations around the world^{184,186}. If the patient cohorts were to be pooled, a larger proportion of SNPs would deviate significantly from Hardy-Weinberg Equilibrium (HWE) and hence be excluded from analysis¹⁸⁵.

The following quality control steps described in the following section are carried out prior to association testing:

2.5.1 Determining level of missingness in the data

For each cohort, individual sample missingness rate is calculated with PLINK. The resulting missingness file for the study cohort is then imported into Microsoft Excel. The mean and standard deviation of the missingness rate is then

calculated for the cohort. Individuals with missingness rates greater than three standard deviations from the mean were removed¹⁸⁵.

2.5.2 Heterozygosity

An excess of autosomal heterozygosity is a sign of possible contamination or poor input DNA quality during genotyping. Reduced heterozygosity is indicative of two individuals that are related to each other¹⁸⁵. Heterozygosity of each individual in each cohort is calculated using PLINK. The output file is then imported into Microsoft Excel to calculate the mean and standard deviation of the *F*-statistic. Individuals with *F*-values less than or more than 3 standard deviations from the mean are removed¹⁸⁷.

2.5.3 Population Stratification

False positives sometimes arise due to differences in minor allele frequencies arising simply by pure chance between cases and controls; the minor allele would be more frequent in either the cases or the controls. In the case of a quantitative trait such as height, the minor allele might be more frequent in the taller individuals than the shorter individuals occurring by pure chance when recruiting them to the study. The lactase gene *LCT* has been demonstrated to be spuriously associated with height in individuals that identify as "white" and or "European descent" that were recruited across different geographical locations across the United States and Europe¹⁸⁴. As such, distance-based clustering methods such as principal components analysis have been developed using the minor allele frequencies of the prospective study cohort to detect and account for such

differences in minor allele frequencies, reducing the likelihood of detecting false positives, allowing the true positives to float to the top¹⁸⁸.

As demonstrated by the study of lactase persistence in European individuals, Price et al.¹⁸⁸ found that genetic variants outside of the *LCT* locus on Chromosome 2 that showed association ($p < 0.05$) with lactase persistence were no longer associated with the phenotype after adjusting for population stratification.

Determining population stratification also helps in the detection of duplicate individuals (same individual genotyped multiple times), as well as closely related individuals in non-family based study designs; closely related individuals can potentially skew the minor allele frequencies and result in false positives or false negatives. The following steps were carried out for the determination of population stratification:

2.5.3.1 Pruning a set of unrelated SNPs

A pruned set of unrelated SNPs is as effective as using the entire genotype dataset in calculating the relatedness of individuals¹⁸⁵. Any two related individuals would have tracts of DNA sharing of approximately 0.6 centimorgans, or approximately 600 kb; a 500 kb pruning window would sufficiently capture the SNPs that are independently occurring pairwise¹⁸⁹. A sliding window step of 50 kb was also selected by taking into account the median haplotype lengths (34.1 kb - 54.4 kb)¹⁵³, such that most often, the independent SNPs would be pruned to be included for calculating the relationship matrix.

2.5.3.2 Calculating the relationship matrix

The pruned set of unrelated SNPs is used to calculate the identity-by-descent/identity-by-state (IBD/IBS) matrix. The purpose of the pruned set of SNPs serves two purposes: a) to remove the stretches of related SNPs, which could result in excessive estimation of relatedness; and b) reducing the amount of computational resource required for the IBS calculation¹⁶⁹.

2.5.3.3 Obtaining the Principal Components Analysis

The IBS/IBD matrix can then be used to obtain the variance-standardized relationship matrix of the unrelated individuals within a study cohort¹⁶⁹. The resultant PCA table expresses the variance of the individuals along continuous axes of variation. The PCA can be used for clustering of individuals into homogeneous groups, determining outliers of genetic ancestry in a study cohort, and for visualisation of the population substructures of individuals in the study.

2.5.3.4 Determining outlying genetic ancestry

The first principal component (PC1) was used to determine outlying genetic ancestry. Such individuals are defined to be six standard deviations away from the mean of the PC1 ($PC1_{mean} \pm 3\sigma$)¹⁸⁸; the calculation step was also carried out in Microsoft Excel.

2.5.3.5 Population Substructure

Population substructure of all individuals from the study can be calculated by first merging the genotype files of all the separate cohorts (DEU, LBN, and

SGP) into a single genotype file. Then, the principal components can be calculated as per [Section 2.5.3.1-2.5.3.3](#) and visualised using Microsoft Excel's scatter plot function.

2.5.3.6 Number of Principal Components to include as covariates in the regression

Calculation of the principal component in [Section 2.5.3.3](#) concomitantly generates eigenvalue tables for plotting scree plots; these show how much of the variation each principal component explains. It is used to determine what is the optimum number of principal components required to adjust for population stratification when doing the regression analysis.

2.5.4 SNP-level QC

2.5.4.1 Analysis of Common Variants - Single Variant Analysis

Hundreds of thousands of variants are genotyped in each GWAS study. Hence there is a need to only include the most informative SNPs that could potentially be true associations. Poorly genotyped SNPs (defined as <95% genotyping rate in cohort) could result in erroneous association and are hence removed.

Hardy-Weinberg Equilibrium (HWE) is the measure between a SNP's expected minor allele frequency and the SNP's actual minor allele frequency as observed in the cohort. A significant deviation from the expected frequency (defined as $p < 1e-5$) indicates that the SNPs is not in HWE. SNPs not in HWE

has the potential to skew association tests which result in erroneous associations, and as such, should be excluded from the analysis¹⁸⁵.

Rare variants ($MAF < 1\%$) require disproportionately large effect sizes for them to be statistically significant even at even modest ($p < 1e-4$) when carrying out single-variant association testing¹⁹⁰. As demonstrated in [Figure 4](#), increases in minor allele frequency increases the statistical power for the detection of true associations.

Figure 4: Graph of Statistical Power and Effect Size

Y-axis is the statistical power to detect an association of a genetic variant. X-axis is the effect size of the genetic variant under the Additive model. This model assumes a sample size of 500, and a significance level of $p < 1e-4$. Diagram calculated using the Genome Association Studies Power Calculator^{191,192}.

2.6 Association Testing

2.6.1 Linear Regression

Association testing using linear regression was carried out in PLINK¹⁷⁰. As the trait (eg. serum endoxifen concentration) is a quantitative trait, the program assumes that each copy of the alternate allele increases or decreases the phenotype mean, in accordance with the regression coefficient.

2.6.2 Visualising the phenotype distribution of our study cohorts

The linear regression model assumes that the quantitative trait is distributed according to the normal distribution. To ascertain the distribution of the measured metabolite concentrations, we plotted histograms with normal curve overlays and performed the Kolmogorov-Smirnov test in R^{166,193}.

2.6.3 Fixed and Random Effects in Meta-Analysis

PLINK employs the inverse variance method using the association effect size (β) and standard errors (SE) calculated for each individual population, adjusted for population stratification; pooling samples would not appropriately account for population stratification.

Results have been summarised under the fixed effects model¹⁹⁴. When the results of fixed effects model differs from the random effects model, it is an indication of population heterogeneity¹⁹⁵ and a prime opportunity to study the underlying genetic architecture differences between these populations.

2.6.4 Visualisation of Association Testing

Association testing results were imported into R^{166,193}. Manhattan and QQ plots were created with the *qqman* package¹⁹⁶.

2.6.5 Genomic Inflation

Genomic inflation is calculated by deriving the median right-tailed χ^2 value from the p value from the list of SNPs tested and dividing it by the median $\chi^2 = 0.456$, under 1 degree of freedom:

$$\lambda = \frac{\chi^2_{median}}{0.456}$$

$\lambda = 1$ indicates no genomic inflation; acceptable range of λ values range from 1 to 1.10¹⁹⁷.

2.7 Haplotype Phasing

Imputation of variants using a reference panel that has been genotyped densely can be used to increase the number of SNPs tested in a sample population that has been genotyped at a subset of SNPs¹⁹⁸. However, phasing of haplotypes need to first be conducted for successful imputation of genotypes. Phasing of haplotypes was done using SHAPEIT¹⁷⁷. We have also created our own custom reference panel by phasing the Singaporean subset of individuals that were genotyped by whole genome sequencing. The custom reference panel was also phased using SHAPEIT¹⁷⁷.

2.8 Genotype Imputation

Imputation of variants was carried out using IMPUTE2¹⁹⁹ with the 1000 Genomes Project (1KGP) reference panels¹⁸⁴, and our own custom Singaporean reference panel for the Singaporean samples.

Genotyping arrays are a cheaper and more efficient method to survey genetic variants across the entire genome. However, the monetary and time savings is at the cost of resolution, because only commonly occurring SNPs are included in the array design²⁰⁰. Hence, when conducting a genome-wide association study, the sentinel variant may not be the causal variant, but is instead another marker that is in linkage disequilibrium with the causal variant.

Whole genome sequencing is currently still too expensive to be done for the hundreds to thousands of individuals in a genome-wide association study. However, Array genotyping in combination with imputation allows for genome-wide association hits to be further fined-mapped, and sometimes to the causal variant.

IMPUTE2 takes the known haplotypes, phased from a genotyping array, and compares it to a reference panel of known haplotypes that have been densely genotyped, and "fills in the gaps" by calculating the probability of occurrences of those missing genotypes based on the haplotypes that are observed during array genotyping (*Figure 5*).

IMPUTE2 also allows for the flexibility of breaking up the genome into segments for parallelisation. Since we have the access to cluster computers and multiple job submissions, massive parallelisation is possible using the bash

scripting method. For our purposes, we broken up the genome into 1 megabase segments when imputing.

Figure 5: Imputation Scenario

A: Study sample is sparsely genotyped, usually from array genotyping platforms. **B:** Study sample is assigned to haplotypes ie phasing. Reference is made to known densely genotyped haplotypes in a reference panel eg. 1000 Genomes Reference Panel. **C:** Inferring from the reference panel, the missing genotypes are filled in to the study sample. Diagram from Li, Y. *et al.* Genotype Imputation. *Annual Review of Genomics and Human Genetics* **10**, 387–406 (2009)¹⁹⁸. Figure reproduced with permission from the publisher.

2.8.1 Imputation Quality Control

Imputation scores range from 0 to 1, with an increasing number indicating greater confidence of correct imputation¹⁹⁹. For the imputed datasets, only SNPs with an imputation information (*INFO*) score of >0.95 , and a threshold score of >0.95 were brought forward. SNPs were then further sifted for completion using the same QC thresholds for SNPs stated in [Section 2.5.4](#).

2.9 Processing of ChIP-seq raw reads from NCBI Sequence Read Archive (SRA)

We processed a subset of chromatin immunoprecipitation sequencing (ChIP-seq) data from GSE155908. The raw reads were downloaded in FASTQ format using the SRA toolkit (version 2.10.8)²⁰¹. Trimming of adapter sequences from sequencing reads was carried out with trim_galore (version 0.6.6)²⁰², using Cutadapt version 1.18²⁰³.

Alignment to the hg19 reference genome was carried out using Burrows-Wheeler Aligner (bwa) (version 0.7.17-r1188)²⁰⁴. SAMtools (version 1.10)²⁰⁵ and the picard toolkit (version 2.24.0)²⁰⁶ were used in the post-alignment processing of the sequence alignment maps.

Peak-calling was carried out using MACS2 (version 2.2.7.1)²⁰⁷. Immunoprecipitated samples were normalised to input controls using deepTools (version 3.5.0)²⁰⁸ and converted to Tiled data files (TDF) with igvtools (version 2.8.13) for visualisation using Integrative Genomics Viewer (IGV)²⁰⁹.

2.10 Statistical Power Calculation

To reject the null hypothesis reliably, there must first be enough statistical power. We use $R^{166,193}$ to calculate the statistical power²¹⁰ ([Table 3](#)):

Table 3: Power Calculation

N	512
α	5e-8
H2	0.18
threshold	29.71
power	0.99
N : number of individuals in the study; α : refers to type-I error threshold, which is set at genome-wide significance threshold. H2 : amount of variance explained; threshold : threshold z-score for given α ; power : number between 0 and 1, defined as the probability of correctly rejecting a null-hypothesis.	

2.11 Interpretation of effect size (β) in the log-transformed linear regression model

The linear regression model of the log-transformed metabolite levels is as follows:

$$\log_{10}(y) = \alpha + \beta x$$

where y is endoxifen concentration, β is the effect size, and x is the number of copies of the alternate allele. α is the y-intercept; it is not used in the rejection of the null-hypothesis in association testing.

To interpret the significance²¹¹ of β with respect to y , the equation is written as:

$$\% \Delta y = 10^{\beta x} \times 100\%$$

A working example is as follows ([Table 4](#)):

Table 4: Interpretation of β with respect to y

β	-0.1961		
Number of alternate alleles	0	1	2
$10^{\beta x} \times 100\%$	100%	63.664891%	40.532183%

2.12 Additional Notes

- All alleles listed are with reference to the plus (+) strand.
- All genomic locations are with reference to version GRCh37 (hg19) of the Human Genome Assembly.

3 Results

3.1 Patient demographic and genetic ancestry

Table 5 tabulates the demographic and clinical characteristics of all the breast cancer patient collections reported in this study. Ancestral principal component analysis (PCA) confirms that all patients from Singapore, Lebanon, and Germany are homogeneous within their own population strata (*Figure 6*).

Population stratification for each individual patient cohort was also determined with PCA. The first two PCs are plotted for each cohort (*Figure 7A-C*). The corresponding scree plots indicating the amount of variance explained by each principal component was also plotted (*Figure 7D-F*).

To minimize the effects of any residual genetic stratification, the top axes of variation in the respective cohorts have been included as covariates in the linear regression model; first 3 PCs for the Singapore cohort, first 5 PCs for the Germany cohort, and the first 10 PCs for the Lebanon cohort (*Figure 7D-F*).

Table 5: Patient Collection Demographics and Characteristics

Breast cancer patient collections			
	SGP (N = 152)	DEU (N = 290)	LBN (N = 70)
Age [median (range)], years	50 (31 - 69)	64 (45 - 82)	43 (24 - 51)
Weight [median (range)], kg	57 (39 - 92)	71 (45 - 144)	69 (41 - 110)
Height [median (range)], cm	156 (134 - 172)	N.A.	162 (146 - 175)
BMI [median (range)], kg/m2	22.8 (14.3 - 38.2)	N.A.	25.8 (15.1 - 40.9)
Menopausal status [N (%)]			
Premenopausal	30 (19.7%)	0 (0%)	70 (100%)
Postmenopausal	122 (80.3%)	286 (98.6%)	0 (0%)
Unknown	0 (0%)	4 (1.4%)	0 (0%)
Prior cancer treatment [N (%)]			
Radiotherapy	101 (66.4%)	262 (90.3%)	60 (85.7%)
Unknown	0 (0%)	2 (0.7%)	0 (0%)
Chemotherapy	124 (81.5%)	65 (22.4%)	61 (87.1%)
Unknown	0 (0%)	3 (1.0%)	0 (0%)
Tumour stage [N (%)]			
I	60 (39.4%)	192 (66.2%)	23 (32.9%)
IIA-B	72 (47.4%)	89 (30.7%)	32 (45.7%)
IIIA-B	17 (11.2%)	7 (2.4%)	15 (21.4%)
IV	3 (2.0%)	0 (0%)	0 (0%)
Unknown	0 (0%)	2 (0.7%)	0 (0%)
Hormone receptor status [N (%)]			
ER+/PR+	123 (80.0%)	248 (85.5%)	55 (78.6%)
ER-/PR+	15 (9.9%)	1 (0.3%)	1 (1.4%)
ER+/PR-	16 (10.5%)	40 (13.8%)	14 (20%)
HER2+	41 (27.0%)	49 (16.9%)	16 (22.9%)
Unknown	0 (0%)	1 (0.3%) unknown (HER2)	0 (0%)

Figure 6: Genetic ancestry of individual cohorts

A-C: Genetic Ancestry of individual cohorts; X-axis is Principal Component 1, y-axis is Principal Component 2. **D-F** Amount of variance explained for each principal component. X-axis is number (#) of Principal Component. Y-axis shows eigenvalue, which is the amount of variance that is explained by each PC. Red horizontal line indicates cut-off number of PCs included the linear regression model.

Figure 7: PCA of genetic ancestry for all patients in the tamoxifen study cohort.

DEU:Germany (Blue); LBN:Lebanon (Orange); SGP:Singapore (Grey). Each point is an individual; Color denotes which population that individual belongs to. Every cohort is clustered within their own strata; there is no overlap of individuals into another strata.

3.2 Data transformation for serum tamoxifen metabolite concentrations

Before conducting the linear regression association testing, a positive skew was observed in the distribution of raw serum endoxifen values in all three of our tamoxifen cohorts (*Figure 8 A, B, and C*). The Kolmogorov-Smirnov (KS) test was also performed for each cohort against a normally distributed string of values with the same mean and standard deviation. Two of the three cohorts, DEU and SGP, demonstrated significant ($p < 0.05$) deviation from the normal distribution (*Figure 8 G, H, and J*). To reduce the effects of skew on the linear regression, we sought to conform the data to a normal distribution by applying a $\log_{10}$ transformation to the endoxifen values. After log-transformation, the skew observed within is noticeably less than before (Data transformation for serum tamoxifen metabolite concentrations *Figure 8 D, E, and F*), and the KS test also demonstrated no deviations from the normal distribution ($p < 0.05$) (*Figure 8 K, L, and M*).

For the other three metabolites (*Figure 9, Figure 10, and Figure 11*) the distributions of the metabolite concentrations did not skew significantly, nor was there any significant deviation from the normal distribution. Hence, a log-transformation of the data points was not necessary but was performed for the purposes of comparing the effect sizes (β) of the linear regression with endoxifen.

Figure 8: Histogram and Kolmogorov-Smirnov test with normal curve overlays for endoxifen

A, B, and C: Histogram of endoxifen values with normal curve overlay in red; **D, E, and F:** Histogram of $\log_{10}$ -transformed endoxifen values with normal curve overlay in blue; **G, H, and J:** Two-sample KS test. Each point is an individual in the respective tamoxifen cohort, ranked in order of increasing raw endoxifen value, red line is the normal curve; **K, L, and M:** Two-sample KS test. Each point is an individual in the respective tamoxifen cohort, ranked in order of increasing $\log_{10}(\text{endoxifen})$ value, blue line is the normal curve.

Figure 9: Histogram and Kolmogorov-Smirnov test with normal curve overlays for 4-hydroxytamoxifen

A, B, and C: Histogram of 4-hydroxytamoxifen values with normal curve overlay in red; **D, E, and F:** Histogram of log-transformed 4-hydroxytamoxifen values with normal curve overlay in blue; **G, H, and I:** Two-sample KS test. Each point is an individual, red line is the normal curve; **J, K, L, and M:** Two-sample KS test. Each point is an individual, blue line is the normal curve.

Figure 10 Histogram and Kolmogorov-Smirnov test with normal curve overlays for *N*-desmethyltamoxifen

A, B, and C: Histogram of *N*-desmethyltamoxifen values with normal curve overlay in red; **D, E, and F:** Histogram of log-transformed *n*-desmethyltamoxifen values with normal curve overlay in blue; **G, H, and J:** Two-sample KS test. Each point is an individual, red line is the normal curve; **K, L, and M:** Two-sample KS test. Each point is an individual, blue line is the normal curve.

Figure 11: Histogram and Kolmogorov-Smirnov test with normal curve overlays for tamoxifen

A, B, and C: Histogram of tamoxifen values with normal curve overlay in red; **D, E, and F:** Histogram of log-transformed tamoxifen values with normal curve overlay in blue; **G, H, and J:** Two-sample KS test. Each point is an individual, red line is the normal curve; **K, L, and M:** Two-sample KS test. Each point is an individual, blue line is the normal curve.

3.3 Association testing of *CYP2D6* Genetic Variants with tamoxifen metabolites

There are more than a hundred different *CYP2D6* alleles discovered to date^{15,134–136}. We conducted association testing for the classical *CYP2D6* alleles by genotyping the allele's corresponding proxy SNP (detailed in [Table 1](#)) in all three of our study cohorts. These SNPs were not captured in the Illumina OmniExpress array, hence they were genotyped using the TaqMan SNP genotyping assay. As we have done in the genome-wide association study, genetic ancestry was included as covariates in the linear regression.

*CYP2D6*4* was strongly associated with decreased serum endoxifen in both the DEU ($\beta = -0.1961, p = 1.095 \times 10^{-13}$) and the LBN ($\beta = -0.2348, p = 2 \times 10^{-4}$) cohorts. However, *CYP2D6*4* was too rare in the SGP cohort ($MAF_{SGP} < 0.01$) to conduct any meaningful association testing ([Table 6](#)).

Here, β indicates the effect size for the linear regression model. Using the example of *CYP2D6*4* in DEU, and a β of -0.1961, carrying one copy of *CYP2D6*4* allele (i.e. heterozygotes) will decrease mean serum endoxifen levels to approximately 63% of the mean of the wild type. Carrying two *CYP2D6*4* alleles further decreases serum endoxifen levels to approximately 40% of the wild type (see [Section 2.11](#) and [Table 4](#)).

*CYP2D6*10* was also strongly associated with decreased serum endoxifen in the Singaporean cohort ($\beta = -0.1485, p = 8.73 \times 10^{-9}$),

surpassing the genome-wide significance threshold. In the German cohort, *CYP2D6*10* is not associated with serum endoxifen ($\beta = -0.1243, p = 0.08$). *CYP2D6*10* is notably missing from the Lebanese cohort, and also is relatively rare in the German cohort ($MAF_{DEU} = 0.02$)

*CYP2D6*41* was the most common variant in the Lebanese cohort, however the association between the variant and serum endoxifen was not even nominally significant ($\beta = -0.08, p = 0.1$). In the Singaporean cohort, *CYP2D6*41* was also not associated with serum endoxifen levels ($\beta = -0.08, p = 0.2$). In the German cohort, *CYP2D6*41* was strongly associated with decreased serum endoxifen ($\beta = -0.21, p = 6.8 \times 10^{-8}$).

In the overall meta-analysis, *CYP2D6*4*, *CYP2D6*10*, and *CYP2D6*41* all surpass the genome-wide significance threshold of ($p < 5 \times 10^{-8}$) (refer to [Table 6](#)). As observed from the Forest plots ([Figure 12](#)), *CYP2D6*4* has the strongest effects, with none of the populations' confidence intervals crossing the null (dotted) line. However, *CYP2D6*4* is also not present from the SGP cohort. *CYP2D6*10* has effects that are less strong than *CYP2D6*4*, but its effects are heterogeneous, with the association testing in DEU having its confidence intervals overlapping the null line. Also, *CYP2D6*10* is not present in the LBN cohort. *CYP2D6*41* has strong, but heterogeneous effects in the cohorts that I have studied. In LBN and SGP, the confidence intervals overlap the null line. LBN is also the cohort where *CYP2D6*41* has the highest minor allele frequency.

Figure 12 summarizes the frustrations of pharmacogenetic testing for *CYP2D6* alleles: the combination of differences in minor allele frequency, together with the strength of association with steady state serum endoxifen levels makes pharmacogenetic predictive testing for *CYP2D6* alleles a tricky business. A panel of between 7 to 11 SNPs need to be tested for, due to minor allele frequency differences. However, the utility for genotyping some alleles in certain populations (eg *CYP2D6*10* in DEU, or *CYP2D6*4* in SGP) is negligible. The other *CYP2D6* variants are also comparatively rare (**Table 7**) and hence provide very little information for the cost of genotyping.

Figure 12: Forest plots of the three most common CYP2D6 variants in our study cohorts

Dotted line indicates the null line; box indicates effect, size of box indicates weight in meta-analysis, horizontal lines indicate 95% confidence interval; effects on the left side of the null line indicate decreasing endoxifen levels per copy of the minor allele, effects on the right side of the null line indicate increasing endoxifen levels per copy of the minor allele. All three CYP2D6 variants indicate decreasing serum endoxifen, consistent with the literature annotations of "non-functioning" (*CYP2D6*4*;A) and "partially functioning" (*CYP2D6*10*;B, *CYP2D6*41*;C)

Table 6: CYP2D6 alleles association testing results in our study cohorts

Allele Name	SNP	Ref ⁸	Alt ⁹	DEU			LBN			SGP			Meta		
				Alt Freq ¹⁰	β	<i>P</i>	Alt Freq	β	<i>P</i>	Alt Freq	β	<i>P</i>	β	<i>P</i>	<i>I</i> ²
CYP2D6*4	rs3892097	C	T	0.19	-0.1961	1.1E-13	0.12	-0.2348	2.04E-04	<0.01	-	-	-0.202	2.28E-18	0
CYP2D6*10	rs1065852	G	A	0.02	-0.1243	0.08	0	-	-	0.57	-0.1485	8.73E-09	-0.146	2.26E-10	0
CYP2D6*41	rs28371725	C	T	0.08	-0.2103	6.82E-08	0.16	-0.08897	0.1558	0.03	-0.08629	0.24	-0.162	4.1E-08	51.13
CYP2D6*5	deletion	WT	del	0.03	-0.1212	0.03	0.04	-0.01267	0.919	0.09	-0.09009	0.12	-0.0964	0.01	0

Table 7: Association testing results of rare CYP2D6 variants

Allele Name	SNP	Ref	Alt	DEU			LBN			SGP		
				Alt Freq	β	<i>P</i>	Alt Freq	β	<i>P</i>	Alt Freq	β	<i>P</i>
CYP2D6*6	rs5030655	A	Del	0.02	-0.13	0.1	<0.01	-	-	<0.01	-	-
CYP2D6*9	rs5030656	TTC	Del	0.03	0.02	0.6	0	-	-	0	-	-
CYP2D6*15	-		137insT	0	-	-	<0.01	-	-	0	-	-
CYP2D6*17	rs28371706	A	G	0	-	-	<0.01	-	-	0	-	-
CYP2D6*29	rs61736512	C	T	0	-	-	0	-	-	0	-	-

⁸ Reference (Ref) in our use refers to the reference allele which is found in the 1000 Genomes Project reference panel

⁹ Alternate (Alt) refers to the alternate allele which is found in the 1000 Genomes Project reference panel.

¹⁰ Alternate allele frequency (Alt freq) refers to the allele frequency of the alternate allele, as described in footnote 9.

3.4 Effectiveness of *CYP2D6* variants at predicting endoxifen levels

*CYP2D6*4* and *CYP2D6*10* are effective predictor variants for the purposes of predicting metaboliser status in their respective cohorts, with R^2 of 0.18 and 0.21 respectively. However, despite *CYP2D6*41* has an equivalent R^2 of 0.17, it is a suboptimal candidate for the same purposes as the association is not significant ($P = 0.1$), indicating that the association could be present due to chance alone ([Figure 13](#), [Table 6](#)). The association for *CYP2D6*41* in LBN could also be influenced by its small sample size, with only 70 patients included. The association of *CYP2D6*41* was significant ($p = 6.82 \times 10^{-8}$) in the DEU cohort, with 290 patients ([Table 6](#)).

Figure 13: Linear regression of most common CYP2D6 variants in study cohorts

Linear regression of *CYP2D6*4* in DEU (A), *CYP2D6*41* in LBN (B), and *CYP2D6*10* in SGP (C). Every point on the graph is an individual in the study cohort. β is the gradient of the regression line; P is statistical significance; R^2 is the coefficient of determination, which is the amount of variance the regression line explains.

3.5 Effectiveness of Activity Score

In the section above, I have demonstrated the individual SNPs that serve as proxies for the *CYP2D6* alleles are strongly associated with endoxifen. I have also determined the amount of variance that each *CYP2D6* allele explained.

Earlier association studies between *CYP2D6* alleles and tamoxifen metabolite concentrations have utilised the Activity Score (AS) method (see Table 2) to capture information on the whole *CYP2D6* gene as a unit. The AS method has also been used to determine how much of the variance was observed in tamoxifen metabolite levels in patient cohorts.

To assign Activity Scores (AS), I used the common *CYP2D6* alleles *CYP2D6*4*, *CYP2D6*5*, *CYP2D6*10*, and *CYP2D6*41* that were individually significantly associated with serum endoxifen levels (presented in Table 6). We used the 2020 guidelines by PharmGKB for *CYP2D6*²¹² for each allele's activity score: *CYP2D6*4* and *CYP2D6*5* has AS = 0; *CYP2D6*10* has AS = 0.25; *CYP2D6*41* has AS = 0.5 (Also summarised in Table 2). We assumed an additive effect in individuals where more than two alleles have been detected during genotyping. Because amplified alleles (eg. *CYP2D6*1xN*) are not included in this regression model, an individual can have the maximum AS of 2, and a minimum AS of 0; the count for each *CYP2D6* diplotype and its corresponding AS is recorded in Table 8. The AS for each cohort was then used to conduct a linear regression with log-transformed endoxifen.

As expected, the AS was strongly associated with log-transformed serum endoxifen levels ($\beta_{meta} = 0.2264$, $p_{meta} = 3.65 \times 10^{-45}$). AS explained 23% of the variance in SGP, 31% in DEU, and 26% in LBN (Table 9, Figure 14).

Although the amount of variance explained is lower than those reported in previous studies^{51,54,56,57,140} (also see [Section 1.8](#)), our regression model only includes 4 *CYP2D6* alleles with no other covariates, in comparison to the 7-11 *CYP2D6* alleles, age, BMI and other genetic covariates (eg *CYP2C9*, *CYP2C19* alleles) included in other studies. Although a direct comparison is not possible with the other studies due to differences in regression parameters, we can still conclude that a large proportion of the variance is explained through the common *CYP2D6* alleles.

Table 8: Summary of CYP2D6 diplotype and Activity Score and counts in each cohort.

Unique Diplotypes	Activity Score	SGP		DEU		LBN	
		Count	Freq	Count	Freq	Count	Freq
<i>CYP2D6*1/*1</i>	2	26	0.17	136	0.47	34	0.49
<i>CYP2D6*1/*4</i>	1	0	0.00	70	0.24	8	0.11
<i>CYP2D6*1/*5</i>	1	10	0.07	11	0.04	4	0.06
<i>CYP2D6*1/*10</i>	1.25	44	0.29	8	0.03	0	0.00
<i>CYP2D6*1/*41</i>	1.5	2	0.01	27	0.09	15	0.21
<i>CYP2D6*4/*4</i>	0	0	0.00	10	0.03	2	0.03
<i>CYP2D6*4/*5</i>	0	0	0.00	1	0.00	1	0.01
<i>CYP2D6*4/*10</i>	0.25	0	0.00	4	0.01	1	0.01
<i>CYP2D6*4+*10/*10</i>	0	2	0.01	0	0.00	0	0.00
<i>CYP2D6*4/*41</i>	0.5	0	0.00	17	0.06	3	0.04
<i>CYP2D6*5/*10</i>	0.25	1	0.01	0	0.00	0	0.00
<i>CYP2D6*5/*10x2</i>	0	14	0.09	0	0.00	0	0.00
<i>CYP2D6*5/*41</i>	0.5	1	0.01	1	0.00	0	0.00
<i>CYP2D6*5/*5</i>	0	0	0.00	2	0.01	0	0.00
<i>CYP2D6*10/*10</i>	0.5	46	0.30	0	0.00	0	0.00
<i>CYP2D6*10/*41</i>	0.75	4	0.03	2	0.01	0	0.00
<i>CYP2D6*41/*41</i>	1	1	0.01	1	0.00	2	0.03
Missing	NA	1	0.01	0	0.00	0	0.00

Table 9: Linear Regression Testing with Activity Score

Cohort	β	SE	T	P	R^2	β_{meta}	SE _{meta}	P_{meta}
DEU	0.2352	0.0206	11.4325	3.31E-25	0.31	0.2264	0.0160	3.67E-45
LBN	0.2313	0.0476	4.8605	7.39E-06	0.26			
SGP	0.2051	0.0305	6.7252	3.57E-10	0.23			
β : Effect size for each activity score count; SE: standard error of linear regression; T: Student's T-statistic; P : p -value; R^2 : Amount of variance explained by regression line.

Figure 14: Linear Regression of Activity Score against log₁₀(Endoxifen)

Each point is an individual in the respective cohorts. Blue lines indicate regression line. R^2 indicates how much of the variance of the log-transformed endoxifen is explained by the regression line.

3.6 GWAS of tamoxifen metabolites

3.6.1 GWAS of endoxifen

In the discovery cohort consisting of 152 individuals from Singapore, we performed a genome-wide association test on steady state serum endoxifen levels. A total of 568,768 SNPs passed stringent quality control and were tested for association with endoxifen. At the suggestive p-value cut-off, there were 41 SNPs, 15 of which surpassed the genome-wide significance threshold. Of the 41 suggestive SNPs, 34 SNPs were found on Chromosome 22; all 15 SNPs with ($p < 5 \times 10^{-8}$) were found on Chromosome 22 (*Figure 15A*).

Considering the findings in the Singapore cohort, we sought to replicate the association in other patient cohorts. Another two patient cohorts were recruited in Germany and in Lebanon, with 290 patients and 70 patients, respectively. Similarly, we performed a genome-wide association test on endoxifen in these two cohorts. In the German cohort, 28 SNPs surpassed the suggestive cut-off threshold, with 22 of those SNPs found on Chromosome 22. 14 SNPs surpassed the genome-wide significance threshold, and all were found on Chromosome 22 (*Figure 15B*). In the Lebanese cohort, 8 SNPs were found to have suggestive significance. Of note, no variants on Chromosome 22 demonstrated even nominal significance (*Figure 15C*).

In all three cohorts, we did not observe any evidence of genomic inflation ($\lambda_{SGP} = 1.01$; $\lambda_{DEU} = 1.00$; $\lambda_{LBN} = 1.00$; see also *Figure 16*) indicating that the association results were not confounded by cryptic population structure.

Figure 15: Manhattan plot of individual cohort endoxifen GWAS

Every point is a SNP in the analysis. On the x-axis is the chromosome number, along x-axis is the chromosome position, increasing from left to right. Black points indicates odd-number chromosomes. Grey points indicates even number chromosomes. On the y-axis is $-\log_{10}(p)$; the smaller the p-value, the larger value. Most significant associations are highest on the Manhattan plot. Blue horizontal line is the suggestive threshold of ($p < 1e-5$). Red horizontal line is the genome-wide significance threshold of ($p < 5e-8$). **A**: Manhattan plot for GWAS of endoxifen in Singapore cohort. **B**: Manhattan plot for GWAS of endoxifen in Germany cohort. **C**: Manhattan plot for GWAS of endoxifen in Lebanon cohort.

Figure 16 Quantile-quantile (QQ) plot of p-values for individual breast cancer patient cohorts

Every point is a SNP in the analysis. On the x-axis is *expected $-\log_{10}(p)$* , on the y-axis is *observed $-\log_{10}(p)$* . Red line is the null line of *observed = expected*. λ is genomic inflation factor. **A:** QQ plot for GWAS of endoxifen in Singapore cohort. **B:** QQ plot for GWAS of endoxifen in Germany cohort. **C:** QQ plot for GWAS of endoxifen in Lebanon cohort.

3.6.2 GWAS meta-analyses of endoxifen in breast cancer patient cohorts

The genome-wide association results from the three cohorts were combined in an inverse-variance meta-analysis, to find genetic variants that are generalised, robustly significant, and commonly shared throughout all populations, of which a total of 627,155 variants were present in the meta-analysis. 546,773 variants were present in all three cohorts. Amongst them, 343,268 variants demonstrated homogeneous effects (which is defined as having a null heterogeneity index score $I^2 = 0$). Amongst the genetic variants with homogeneous effects, 14 SNPs surpassed the suggestive threshold, with 11 SNPs found on Chromosome 22. 10 SNPs surpassed the genome-wide significance threshold, and all were found on Chromosome 22 ([Figure 17](#)).

The strongest association was observed at rs5751247 ($p_{meta} = 8.98 \times 10^{-28}$), located ~21 kb upstream of transcription factor 20 (*TCF20*) ([Figure 18Error! Reference source not found.](#)), also ~106 kb upstream of *CYP2D6*. Of note, there are no genetic markers represented in the *CYP2D6* gene itself.

Figure 17: Manhattan plot of endoxifen meta-analysis

Every point is a SNP in the analysis. On the x-axis is the chromosome number, along x-axis is the chromosome position, increasing from left to right. Red indicates odd-number chromosomes. Blue indicates even number chromosomes. On the y-axis is $-\log_{10}(p)$; the smaller the p-value, the larger value. Most significant associations are highest on the manhattan plot. Blue horizontal line is the suggestive threshold of ($p < 1e-5$). Red horizontal line is the genome-wide significance threshold of ($p < 5e-8$). When plotting the manhattan plots for the meta-analysis, only SNPs that were present in all three cohorts (DEU, LBN, and SGP) with a heterogeneity score of zero ($I^2 = 0$) are included in the plot.

Figure 18:Regional association for Chromosome 22 locus in the meta-analysis of log-transformed Endoxifen levels
 Genetic marker rs5751247 demonstrates the strongest evidence of statistical association in the locus. No genetic markers in CYP2D6 genotyped on the GWAS chip.

3.6.3 Meta-analysis of GWAS on other tamoxifen metabolites

We also conducted GWAS on the steady state amounts of the other two main tamoxifen metabolites, NDM and 4OHT, as well as tamoxifen itself.

Nine variants were genome-wide significant in the NDM analysis but only 4 demonstrated homogeneous effects and all were found on Chromosome 22 (*Figure 19A*). rs5751247 was also found to be genome-wide significant but did demonstrate some degree of heterogeneity ($\beta = 0.0733, p = 7.5 \times 10^{-11}, I^2 = 32.47$).

In the overall meta-analyses for 4OHT and tamoxifen, no markers were found to surpass the genome-wide significance threshold (*Figure 19B* and *Figure 19C* respectively), suggesting that genetic control for 4OHT and tamoxifen is not significant within our experimental context.

rs5751247 was found to surpass the suggestive significance threshold in the 4OHT analysis ($\beta = -0.0618, p = 5.8 \times 10^{-07}, I^2 = 0$), which indicates that perhaps with a larger sample size, this association could surpass the genome-wide association threshold.

No evidence of genomic inflation was found in the association tests for NDM, 4OHT, or tamoxifen (λ_{GC} between 1.00-1.02 in all analyses) (*Figure 20*).

Figure 19: Manhattan plot of meta-analyses for N-desmethyltamoxifen, 4-hydroxytamoxifen, and Tamoxifen

All SNPs included in above Manhattan plots were present and homogeneous in all three breast cancer patient cohorts in their respective analyses. **A:** N-desmethyltamoxifen meta-analysis. **B:** 4-hydroxytamoxifen meta-analysis. **C:** Tamoxifen meta-analysis.

Figure 20: QQ plot for all tamoxifen metabolite GWAS meta-analyses

Every point is a SNP in the analysis. On the x-axis is *expected* $-\log_{10}(p)$, on the y-axis is *observed* $-\log_{10}(p)$. Red line is the null line of *observed* = *expected*. λ is genomic inflation factor. No genomic inflation observed in all metabolite GWAS analyses.

3.7 Fine mapping reveals index SNP rs5751245

We carried out imputation in the genomic region of Chromosome 22:41Mb-44Mb to fine-map the association of genetic variants and endoxifen. Within this loci contains the *CYP2D6* gene that has been known to affect the metabolism of tamoxifen to endoxifen⁴⁶.

With the Illumina OmniExpress Beadchip platform, only a subset of SNPs has been directly genotyped; of note, no genetic variants within the *CYP2D6* gene itself was genotyped using the Beadchip method. Imputation using a densely genotyped reference panel (whole-genome sequencing dataset and 1000 Genomes Reference Panel) increases the number of SNPs we can test for association in the particular loci in hopes of finding the causal genetic variant.

After imputation, a total of 5,179 SNPs passed quality control measures (refer to [Section 2.8](#)) and were tested for association with steady state log-transformed serum endoxifen and NDM levels. We identified a strong signal at rs5751245 ($\beta = -0.1739, p = 1.2 \times 10^{-28}, I^2 = 0$) in our endoxifen analysis. In the NDM analysis, rs5751245 is also the most significant association ($\beta = -0.0754, p = 2.07 \times 10^{-11}, I^2 = 25.81$) with a moderate amount of heterogeneity observed in the meta-analysis. rs5751245 is annotated as an upstream variant of *TCF20*, and approximately 94kb upstream of *CYP2D6* ([Figure 21](#)).

Figure 21 Regional Manhattan plot of Chromosome 22:41Mb-44Mb

Regional association plot of rs5751245. Colour scheme for variants in the plot indicate linkage disequilibrium with rs5751245, according to the 1KGP_EUR cohort. Blue vertical bar marks genomic location of rs5751245; Red vertical bar indicates genomic location of CYP2D6.

3.7.1 Effectiveness of rs5751245 at explaining variance in comparison to *CYP2D6* alleles

To evaluate the viability of rs5751245 as a proxy SNP for predicting endoxifen, we compared the variance explained (R^2 of the linear regression) of rs5751245 to the variance explained of the most common *CYP2D6* alleles in each study cohort. rs5751245 explained 18-22% of the variance observed in log-transformed endoxifen levels. There is also sufficient power in our study to reject the null hypothesis (Power=0.99, [Table 3](#))

rs5751245 is more effective than *CYP2D6*4* in the DEU cohort, it is better than *CYP2D6*41* (moderate P-value versus null) in the LBN cohort, and is slightly worse than *CYP2D6*10* ([Figure 22](#) and [Table 10](#)). As a whole, rs5751245 is an appropriate proxy to replace the most common *CYP2D6* alleles found in the Caucasian, Middle-Eastern Arab, and East Asian populations.

Figure 22: rs5751245 linear regression plots

Variance explained for each cohort summarised in [Table 10](#). Each individual is represented by a point on the graph. Red lines indicate linear regression line for each of the cohort studied. Y-axes indicated $\log_{10}(\text{endoxifen})$ concentration. X-axes indicate number of copies of the alternate allele of rs5751245 is present.

Table 10: Comparison between rs5751245 and most common *CYP2D6* allele

	DEU rs3892097 (<i>CYP2D6</i> *4)			LBN rs28371725 (<i>CYP2D6</i> *41)			SGP rs1065852 (<i>CYP2D6</i> *10)		
	R^2	MAF	P	R^2	MAF	P	R^2	MAF	P
rs5751245	0.22	0.2927	8.468E-17	0.18	0.2246	0.001	0.18	0.4079	4.395E-8
Most common <i>CYP2D6</i> allele	0.18	0.1931	1.1E-13	0.17	0.1594	0.1557	0.21	0.57	8.73E-9
R^2 : variance explained by regression model. MAF: Minor allele frequency									

3.7.2 Conditional analysis on index SNP rs5751245

rs5751245 is the sole index SNP; conditional analysis on rs5751245 allele dosage did not reveal a second index SNP at this particular locus ([Figure 23](#)). As none of the remaining SNPs are statistically associated with the trait, this tells us that the other SNPs that were also associated with endoxifen are in fact in linkage disequilibrium with rs5751245.

Figure 23: Regional Manhattan plot of Chromosome 22:41Mb-44Mb; Conditional analysis on rs5751245

A: Regional association plot for endoxifen conditioned on rs5751245 allele dosage. **B:** Regional association plot for *N*-desmethyltamoxifen conditioned on rs5751245 allele dosage. None of the remaining SNPs are genome-wide significant.

3.8 Endoxifen and *N*-desmethyldoxifen SNPs

Of the four compounds that we conducted GWAS for, only endoxifen and NDM had SNPs that surpassed genome-wide significance in our experimental context (see [Figure 17](#) and [Figure 19](#)). We next looked at SNPs that were common in both our endoxifen and NDM analyses to understand the relationship between the SNPs shared by endoxifen and NDM.

We looked for SNPs that were genome-wide significant ($p < 5 \times 10^{-8}$) and demonstrated homogeneous effects across all three ethnic cohorts ($I^2 = 0$) in both our endoxifen and NDM analyses. 20 such SNPs were shared between the endoxifen and the NDM analyses ([Table 11](#)). Interestingly, these 20 SNPs have their effect sizes in the opposite direction; SNPs that increase endoxifen levels per copy of the alternate allele decreases NDM levels, and vice versa. Increased formation of endoxifen leads to lesser NDM at steady state, and vice versa.

The association testing results in [Table 11](#) tells us two things. Firstly, it supports the biochemical evidence of NDM as a precursor of endoxifen⁴⁶. Secondly, it demonstrates the lack of utility for measuring both NDM and endoxifen, since the endoxifen measurement would indirectly inform the NDM measurement. Since endoxifen is the therapeutic compound and NDM demonstrates no therapeutic effect, the measurement and prediction of endoxifen is the one that provides utility.

Table 11: SNPs with no heterogeneity ($I^2 = 0$) in both endoxifen and NDM analyses

				ENDO			NDM		
Chr	BP	SNP	Alt Allele	<i>P</i>	β	I^2	<i>P</i>	β	I^2
22	42465788	rs133373	A	3.09E-21	0.1416	0	3.47E-09	-0.0615	0
22	42466905	rs133376	T	6.49E-21	0.1404	0	3.40E-09	-0.0615	0
22	42473603	rs133383	T	1.12E-20	0.1399	0	1.55E-09	-0.0629	0
22	42481400	rs5758574	C	5.71E-22	-0.145	0	7.42E-10	0.0634	0
22	42481849	rs7245	A	8.73E-22	-0.1423	0	2.52E-09	0.0617	0
22	42485671	rs2284087	C	1.76E-22	-0.1442	0	3.24E-10	0.0645	0
22	42489434	rs5758577	C	3.75E-22	-0.1432	0	1.77E-09	0.0622	0
22	42490292	rs5758578	C	1.84E-22	-0.1439	0	2.50E-09	0.0615	0
22	42496506	rs6002607	T	1.93E-22	-0.145	0	4.99E-09	0.0605	0
22	42496629	rs6002608	C	2.32E-22	-0.1453	0	4.92E-09	0.0608	0
22	42505841	rs2413668	A	2.56E-23	-0.1463	0	4.99E-09	0.06	0
22	42508937	rs5751216	T	1.57E-23	-0.1473	0	1.08E-09	0.0624	0
22	42509666	rs6002620	G	1.87E-23	-0.1474	0	1.06E-09	0.0626	0
22	42511822	rs2413670	C	1.21E-23	-0.1474	0	1.05E-09	0.0624	0
22	42511825	rs2413671	A	1.21E-23	-0.1474	0	1.05E-09	0.0624	0
22	42516205	rs5751220	C	6.98E-23	-0.146	0	3.15E-09	0.0616	0
22	42516366	rs5751221	T	3.08E-24	-0.1706	0	3.04E-08	0.0676	0
22	42517636	rs5758587	G	3.47E-23	-0.1463	0	3.81E-09	0.061	0
22	42517989	rs6002626	G	1.60E-22	-0.1447	0	4.58E-09	0.0609	0
22	42518382	rs5758589	G	4.01E-23	-0.1459	0	1.09E-09	0.0626	0
Orange cells indicate positive effect size (β); blue cells indicate negative β									

3.9 Conditional analysis on *CYP2D6**4, *CYP2D6**10, and *CYP2D6**41 diplotype

Specifically, when we adjusted for the allele dosages of the three classical *CYP2D6* alleles (rs3892097 for *CYP2D6**4, rs1065852 for *CYP2D6**10, and rs28371725 for *CYP2D6**41) that we earlier found to be significantly associated with endoxifen, the association signal for all other SNPs were extinguished, including rs5751245 ($\beta = -0.0256$, $p = 0.3$). This analysis suggests that the index SNP rs5751245 explains as much of the variance observed as the three most common and most significantly associated *CYP2D6* variants. In addition, conditional analysis on the three *CYP2D6* allele proxy SNPs revealed the association of *CYP2D6**5 with endoxifen ($\beta_{cond} = -0.145$, $p_{cond} = 1.06 \times 10^{-5}$) ([Table 12](#), and [Figure 24](#)).

*CYP2D6**5 denotes the deletion of the whole *CYP2D6* gene on one of the chromosomes and is a comparatively rare variant (see MAF in [Table 6](#)). The rise in significance of *CYP2D6**5 when conditioned on the three *CYP2D6* allele SNPs with genome-wide significance gives genetic evidence of *CYP2D6*'s definitive role in the metabolism of endoxifen; in the absence of other more commonly occurring genetic variants of strong effect, a rarer variant with equally strong effects becomes apparent.

Figure 24: Regional association plot of endoxifen conditioned on *CYP2D6*4*, *CYP2D6*10*, and *CYP2D6*41* diplotype

Green point marks association testing of *CYP2D6*5* allele. Blue horizontal line indicates suggestive significance threshold ($p < 1e-5$)

Table 12: *CYP2D6*5* endoxifen association testing results

Test	β	P
No condition	-0.096	0.01
Condition rs5751245	-0.128	2.51E-04
Condition <i>CYP2D6*4</i> , <i>CYP2D6*10</i> , and <i>CYP2D6*41</i> Diplotype	-0.145	1.06E-05

3.10 4-hydroxytamoxifen (4OHT) Association Testing

With 4OHT, the index SNP was rs9607885 ($\beta = -0.071$, $p = 7.69 \times 10^{-8}$) (*Error! Reference source not found.A*). Although it does not surpass the genome-wide significance threshold, the results suggest the presence of genomic control on steady-state 4OHT metabolism. rs9607885 is also in strong linkage disequilibrium with the endoxifen index SNP rs5751245 (Pairwise R^2 in SGP = 0.9; DEU = 0.6; LBN = 0.3). Due to the high degree of linkage disequilibrium between the sentinel SNPs of 4OHT and endoxifen, it is highly likely they belong to the same genetic signal due to linkage disequilibrium. Indeed, upon conditional analysis on the allele dosage of rs5751245, the association of the remaining SNPs with 4OHT was abolished (*Figure 25B*). The same genetic variant(s) responsible for variation of steady state serum endoxifen levels is also responsible to 4OHT levels.

A

B

Figure 25: Regional association plot of Chromosome 22:41Mb-44Mb for 4OHT

A) No SNP surpass genome-wide association threshold with 4OHT. Red horizontal lines indicate threshold for genome-wide significance ($p < 5e-8$); Blue horizontal line indicates suggestive significance threshold ($p < 1e-5$). **B)** when conditioned on rs5751245, the other SNPs that were previously associated with 4OHT were abolished.

3.11 GWAS for eQTL of select genes on Chromosome 22

Tamoxifen is metabolised in the liver by cytochrome P450 enzymes to the therapeutically active metabolite, endoxifen⁴⁶. Gene expression is a significant portion of overall enzyme activity as greater levels of transcription leads to higher levels of protein expressed. To explore the role that genetic variation has on gene expression, we conducted a GWAS for expression quantitative trait loci (eQTL) of the genes located in a 2 megabase (Mb) region of rs5751245 (flanking 1 Mb upstream and 1 Mb downstream) on liver tissue. *Figure 26* depicts the genes found in that genomic region. An eQTL refers to a genomic locus that explains variation observed in the amount of mRNA for a gene. In total, there are 57 genes in this 2 Mb region, of which 39 genes' expression was detected through RNA sequencing (RNA-seq) in liver tissue; all 39 genes are protein coding genes.

Figure 26: Regional association plot of rs5751245

On the x-axis is genomic loci Chromosome 22:41.6Mb-43.6Mb; on the y-axis of the Manhattan Plot $-\log_{10}(p)$ for genetic variants associated with endoxifen. Genes found within this loci are plotted underneath the Manhattan plot. Direction of the gene is indicated by the arrowhead, and length of the gene is indicated by the length of the blue bars. Exons are indicated as thick blue bars, and introns are indicated as thin bars. R^2 colour coding depicts the pairwise linkage disequilibrium of other SNPs to the index SNP rs5751245 according to the 1KGP_EUR cohort.

3.11.1 Liver Study (SCRB) Cohort

Our colleagues at the Genome Institute of Singapore Stem Cell and Regenerative Biology (SCRB) group has very kindly shared with us RNA-sequencing and Illumina array genotyping data for us to conduct the eQTL GWAS analysis.

After strict quality controls have been applied to the SCRB cohort, 80 subjects remain in study (48 Chinese, 21 Malays, 11 Indians). All subjects are recruited from Singapore. Ancestral PCA confirms that all patients from the different self-reported ethnicities are homogeneous within their own population strata (*Figure 27A*). Due to the small sample size, GWAS was carried out in this heterogeneous cohort with the inclusion of the first five PCs, although the scree plot demonstrates only three PCs are required (*Figure 27B*). Two additional PCs were included to reduce the effects of cryptic genetic ancestry because of a heterogeneous cohort, whilst at the same time avoid over-fitting of the linear regression model.

Figure 27 Genetic Ancestry Analysis for SCRB Cohort

A: Principal Component Analysis of SCRB cohort PC1-PC2; **B:** Scree plot for SCRB cohort; first five principal components included as covariates in regression analysis.

3.11.2 eQTL GWAS results

A total of 371,227 SNPs passed stringent quality control thresholds and were included in the eQTL GWAS for the 39 genes with detectable expression. Only two of the 39 gene expressions had SNPs surpassing the genome-wide significance threshold of ($p < 5e-8$) ([Figure 28](#)). Those two genes are *CYP2D6* and *NDUFA6* ([Figure 29](#)). We detect no trace of genomic inflation in the eQTL GWAS ($\lambda = 1.03$) ([Figure 30](#)).

The box plot of the gene expression data ([Figure 31](#)) demonstrates the large variance observed in *CYP2D6* expression as well as *NDUFA6* expression compared to the other genes, which supports the GWAS eQTL results observed, that indeed, there is a strong genetic effect exerted on gene expression only on those two particular genes and not others in liver tissue.

Four SNPs were strongly associated with *CYP2D6* eQTL on Chromosome 22 ($p < 5e-8$). Curiously, the same four SNPs were also strongly associated with *NDUFA6* eQTL. However, the sentinel SNP for each gene expression was different; *CYP2D6* eQTL sentinel SNP was rs1135840 ($\beta = 4.168$, $p = 3.9 \times 10^{-15}$) whereas *NDUFA6* eQTL sentinel SNP was rs5751206 ($\beta = 2.275$, $p = 4.194 \times 10^{-15}$) ([Figure 32](#)).

We asked if there could be a common genetic variant that *CYP2D6* and *NDUFA6* eQTLs can potentially share. To answer the above question, we performed imputation for the genomic location of Chromosome 22:41,000,000-44,000,000. After imputation, the index SNP for *CYP2D6* eQTL is rs28371699 ($\beta = 4.6$, $p = 5.9 \times 10^{-18}$) and the index SNP for *NDUFA6* eQTL is

rs133373 ($\beta = 2.4$, $p = 9.7 \times 10^{-17}$) ([Figure 33](#)). In both gene eQTL analyses, imputation uncovered SNPs whose association is stronger than those found on the genome scan chip.

Conditional analysis on the allele dosage of the sentinel SNPs revealed that *CYP2D6* and *NDUFA6* eQTLs are in linkage disequilibrium; the association for rs28371699 with *CYP2D6* eQTL is severely attenuated when conditioned on the allele dosage of rs133373, and the vice versa is also true ([Figure 34](#)).

CYP2D6 shares the same eQTL as *NDUFA6* in the liver; the SNPs which are genome-wide significant for *CYP2D6* are also genome-wide significant for *NDUFA6*. ([Figure 35A](#)). In the 3-dimensional regional association plot where p-values for the *CYP2D6* and *NDUFA6* eQTL are plotted on the y- and z-axis respectively, the SNPs that are strongly associated with the eQTL are located around position Chr22:42.5Mb ([Figure 35B](#)). A total of 64 SNPs surpass the genome-wide significant threshold. Furthermore, all 64 SNPs demonstrate the same direction of effect for both *CYP2D6* and *NDUFA6* gene expression ([Table 13](#))

Figure 28: GWAS of 39 genes eQTL

Manhattan plots of 39 Genes sorted by alphabetical order. Two genes that have SNPs with ($p < 5e-8$); highlighted with black boxes.

Figure 29: Manhattan plots of *CYP2D6* and *NDUFA6* eQTL GWAS

Manhattan plots for *CYP2D6* eQTL GWAS (**A**) and *NDUFA6* eQTL GWAS (**B**). Every point is a SNP in the analysis. On the x-axis is the chromosome number, along x-axis is the chromosome position, increasing from left to right. Red indicates odd-number chromosomes. Blue indicates even number chromosomes. On the y-axis is $-\log_{10}(p)$; the smaller the p-value, the larger value. Most significant associations are highest on the manhattan plot. Blue horizontal line is the suggestive threshold of ($p < 1e-5$). Red horizontal line is the genome-wide significance threshold of ($p < 5e-8$).

Figure 30: QQ plots of *CYP2D6* and *NDUFA6* eQTL GWAS

Every point is a SNP in the analysis. On the x-axis is *expected* $-\log_{10}(p)$, on the y-axis is *observed* $-\log_{10}(p)$. Red line is the null line of *observed* = *expected*. λ is genomic inflation factor. **A**: QQ plot for *CYP2D6* eQTL GWAS. **B**: QQ plot for *NDUFA6* eQTL GWAS. For both genes, most SNPs followed the expected p-values, and only the most significant few SNPs deviated significantly from the expected. This result suggests that the association between the variant and the trait is a true positive association.

Figure 31: Box plot of gene expression of genes in Chr22:41.6-43.6Mb

Y-axis shows $\log_2(\text{TPM})$ normalised gene expression. Genes ranked according to median expression level, from left-to-right high-to-low. "X" within the bar indicates mean expression level. *TPM*: Transcripts per million

Figure 32: Regional association plots of *CYP2D6* and *NDUFA6* eQTL GWAS

Chr22:41Mb-44Mb regional association plots for *CYP2D6* eQTL GWAS (A) and *NDUFA6* eQTL GWAS (B). Each point represents the p-value for the association between a genetic variant and gene expression. Y-axes show $-\log_{10}(p)$ values. Blue horizontal line is the suggestive threshold of ($p < 1e-5$). Red horizontal line is the genome-wide significance threshold of ($p < 5e-8$)

Figure 33: Regional association plots of *CYP2D6* and *NDUF6* eQTL

Regional association plots following genotype imputation of Chr22:41Mb–44Mb for *CYP2D6* eQTL GWAS (A) and *NDUF6* eQTL GWAS (B)

Figure 34: Regional association plots of *CYP2D6* and *NDUF6* eQTL Conditional Analysis

Regional association plots of *CYP2D6* and *NDUF6* eQTL following conditional analysis on allele dosage of index SNPs rs28371699 and rs133373. A: *CYP2D6* eQTL conditioned on rs28371699 allele dosage; B: *NDUF6* eQTL conditioned on rs28371699 allele dosage; C: *CYP2D6* eQTL conditioned on rs133373 allele dosage; D: *NDUF6* eQTL conditioned on rs133373 allele dosage.

Figure 35: Comparison of p-values between *CYP2D6* and *NDUFA6* eQTL GWAS analysis

A: X-axis shows $-\log_{10}(p)$ values in *CYP2D6* eQTL analysis, Y-axis shows $-\log_{10}(p)$ values in *NDUFA6* eQTL analysis. Red and blue lines indicate threshold for genome-wide significance ($p < 5 \times 10^{-8}$); **A** demonstrates that SNPs that are genome-wide significant for one gene eQTL is also genome-wide significant for the other gene eQTL. **B:** X-axis indicates genomic location on Chromosome 22, Y-axis shows $-\log_{10}(p)$ values in *CYP2D6* analysis, Z-axis shows $-\log_{10}(p)$ values in *NDUFA6* analysis; **B** reveals the genomic location of the SNPs to be around Ch22:42.5Mb.

Table 13 :Table of SNPs with $p < 5e-8$ in *CYP2D6* and *NDUFA6* eQTL

Chr	BP	SNP	Alt Allele	CYP2D6 eQTL		NDUFA6 eQTL	
				β	P	β	P
22	rs739294	42381726	C	3.711	4.37E-10	1.96	1.36E-11
22	rs133317	42402386	A	4.012	3.73E-10	2.05	2.31E-11
22	rs34277770	42443908	G	4.062	5.47E-11	2.23	4.2E-14
22	rs10854749	42447341	T	4.062	8.44E-11	2.21	7.21E-14
22	rs133373	42465788	A	4.343	9.26E-13	2.4	9.74E-17
22	rs133376	42466905	T	4.343	9.26E-13	2.4	9.74E-17
22	rs133380	42468910	T	3.927	1.68E-10	2.19	6.07E-14
22	rs133383	42473603	T	4.386	4E-13	2.35	5.85E-16
22	rs5758574	42481400	C	-4.412	1.91E-13	-2.25	1.42E-14
22	rs7245	42481849	A	-4.496	1.78E-14	-2.27	6.06E-15
22	rs2284087	42485671	C	-4.412	1.91E-13	-2.25	1.42E-14
22	rs5758577	42489434	C	-4.479	2.93E-14	-2.28	4.19E-15
22	rs5758578	42490292	C	-4.412	1.91E-13	-2.25	1.42E-14
22	rs6002607	42496506	T	-4.412	1.91E-13	-2.25	1.42E-14
22	rs6002608	42496629	C	-4.412	1.91E-13	-2.25	1.42E-14
22	rs2413668	42505841	A	-4.479	2.93E-14	-2.28	4.19E-15
22	rs5751216	42508937	T	-4.479	2.93E-14	-2.28	4.19E-15
22	rs6002620	42509666	G	-4.412	1.91E-13	-2.25	1.42E-14
22	rs2413670	42511822	C	-4.479	2.93E-14	-2.28	4.19E-15
22	rs2413671	42511825	A	-4.479	2.93E-14	-2.28	4.19E-15
22	rs5751220	42516205	C	-4.442	1.11E-13	-2.24	2.97E-14
22	rs5758587	42517636	G	-4.487	2.72E-14	-2.27	6.32E-15
22	rs6002626	42517989	G	-4.504	1.74E-14	-2.26	9.11E-15
22	rs5758589	42518382	G	-4.421	1.78E-13	-2.25	2.14E-14
22	rs34385013	42521929	A	4.367	3.24E-16	2.04	3.36E-12
22	rs35028622	42522074	A	4.367	3.24E-16	2.04	1.88E-12
22	rs1135840	42522613	C	4.168	3.92E-15	2.04	1.99E-13
22	rs1058164	42525132	C	4.489	3.86E-16	2.14	1.76E-12
22	rs71328650	42525952	A	4.565	9.58E-17	2.12	1.53E-12
22	rs28371699	42526484	C	4.683	5.91E-18	2.15	4.61E-13
22	rs28735595	42528028	T	4.433	3.93E-16	2.03	1.31E-11
22	rs28680494	42528851	C	4.423	1.66E-16	2.02	7.98E-12
22	rs28439297	42528858	T	4.423	1.66E-16	2.02	7.98E-12
22	rs28542726	42529407	T	4.4	1.11E-15	2.27	1.23E-14
22	rs28579115	42530826	T	4.419	3.94E-15	2.1	4.01E-12
22	rs1800754	42537597	G	4.335	1.3E-12	2.24	3.68E-14

22	rs3021083	42538103	C	4.335	1.3E-12	2.24	3.68E-14
22	rs36093924	42538399	T	4.335	1.3E-12	2.24	3.68E-14
22	rs2267448	42538618	G	4.402	3.13E-12	2.26	1.18E-13
22	rs769261	42539623	A	4.275	2.84E-12	2.23	2.98E-14
22	rs769260	42539720	G	4.273	1.76E-12	2.22	4.31E-14
22	rs2743461	42540551	G	4.273	1.76E-12	2.22	4.31E-14
22	rs5758605	42541349	A	4.273	1.76E-12	2.22	4.31E-14
22	rs2142694	42542870	A	4.273	1.76E-12	2.22	4.31E-14
22	rs2743462	42543288	G	4.273	1.76E-12	2.22	4.31E-14
22	rs2743465	42547739	A	4.317	9.11E-13	2.24	2.92E-14
22	rs5758619	42557710	C	4.305	5.82E-13	2.22	4.33E-14
22	rs5758623	42566314	C	4.434	7.94E-12	2.24	1.09E-12
22	rs2413684	42574830	T	4.287	8.93E-13	2.23	3.04E-14
22	rs2899355	42579520	C	4.305	5.82E-13	2.22	4.33E-14
22	rs5758645	42598951	G	4.24	1.68E-12	2.24	4.46E-14
22	rs5751241	42600589	A	4.305	5.82E-13	2.22	4.33E-14
22	rs5758653	42613485	T	4.246	1.82E-12	2.21	7.32E-14
22	rs5758660	42623718	A	4.336	2.04E-12	2.23	9.49E-14
22	rs5751250	42639645	G	4.343	3.54E-12	2.21	1.71E-13
22	rs5758677	42643039	C	4.336	2.04E-12	2.23	9.49E-14
22	rs5751255	42648408	T	4.304	4.38E-12	2.23	1.62E-13
22	rs134869	42652074	A	4.304	4.38E-12	2.23	1.62E-13
22	rs134873	42657566	G	3.81	1.24E-09	2	3.38E-11
22	rs134877	42662371	T	4.218	1.01E-11	2.22	7.76E-13
22	rs134879	42664201	C	4.111	3.98E-11	2.15	1.28E-12
22	rs134882	42670965	C	3.788	3.53E-09	2.02	3.79E-11
22	rs134900	42683343	C	4.069	7.18E-11	2.13	2.3E-12
22	rs134902	42683997	A	4.033	1.44E-10	2.11	4.75E-12
SNPs are sorted in ascending BP. Orange cells indicate $\beta > 0$; blue cells indicate $\beta < 0$. Index SNPs are bolded							

3.12 *CYP2D6* eQTL effect on endoxifen

Next, we investigated the effects of the *CYP2D6* eQTL index SNP on serum endoxifen levels. In the endoxifen analysis, rs28371699 is only strongly associated with endoxifen in our Singaporean cohort ($\beta = 0.1833$, $p = 3.1 \times 10^{-9}$), and marginally associated in the Lebanese cohort ($\beta = 0.09514$, $p = 0.02$); rs28371699 was not imputed in the German cohort.

From the list of 64 SNPs that were genome-wide significant in our eQTL analysis, 23 SNPs also surpassed the genome-wide significance threshold and demonstrates lack of heterogeneity ($I^2 = 0$) in the endoxifen analysis. Of the 23 SNPs, rs2413670 has the strongest association with endoxifen ($\beta_{endo} = -0.1474$, $p_{endo} = 1.2 \times 10^{-23}$). In the eQTL analysis cohort, rs2413670 is in perfect linkage disequilibrium with rs133373 (pairwise $R^2 = 1$) and is in strong linkage disequilibrium with rs28371699 (pairwise $R^2 = 0.85$). Furthermore, according to the 1000 Genomes database, rs2413670 is in strong linkage disequilibrium with rs28371699 in all cohorts (pairwise $R^2 = 0.89$), making it an ideal proxy SNP for *CYP2D6* eQTL. **Table 14** **Error! Reference source not found.** summarises the effects of our sentinel SNPs.

rs133373 is annotated as an intronic variant of *NAGA* and is our index SNP for *NDUFA6* eQTL. However, no SNPs were strongly associated with *NAGA* expression in our eQTL analysis. *NDUFA6-DT* is a long non-coding RNA whose expression was not detected in the RNA sequencing.

Table 14: Summary information of index and proxy SNPs

dbSNPID	Chr	BP	Alt/Ref	β_{CYP2D6}	P_{CYP2D6}	β_{NDUFA6}	P_{NDUFA6}	β_{ENDO}	P_{ENDO}	R^2_{ENDO}	Gene	Variant annotation
rs133373	22	42465788	A/G	4.343	9.20E-13	2.4	9.70E-17	0.1416	3.00E-21	0	<i>NAGA</i>	intron
rs2413670	22	42511822	C/G	-4.479	2.90E-14	-2.28	4.10E-15	-0.1474	1.20E-23	0	<i>NDUFA6-DT</i>	intron
rs28371699	22	42526484	C/A	4.683	5.90E-18	2.15	4.60E-13	0.1547	8.30E-11	66.7	<i>CYP2D6</i>	intron
rs5751245	22	42620940	C/T	-2.702	0.0005	-1.107	0.002	-0.1739	1.20E-28	0	<i>TCF20</i>	intron
SNPs are sorted in ascending BP. Orange cells indicate $\beta > 0$; blue cells indicate $\beta < 0$. Index SNPs are bolded												

3.12.1 eQTL analysis conditional on rs5751245 allele dosage

The index SNP for endoxifen rs5751245 is not associated with CYP2D6 or NDUFA6 eQTL ([Table 14Error! Reference source not found.](#)). In the eQTL conditional analysis on the allele dosage of rs5751245, no significant change was observed. Very slight attenuation was observed in our index SNPs rs28371699 and rs133373 ([Figure 36](#)).

Figure 36: Regional association plots of CYP2D6 and NDUFA6 eQTL Conditional Analysis

Regional association plots of *CYP2D6* and *NDUFA6* eQTL following conditional analysis on allele dosage of endoxifen index SNPs rs5751245. **A:** *CYP2D6* eQTL conditioned on rs5751245 allele dosage; **B:** *NDUFA6* eQTL conditioned on rs5751245 allele dosage.

3.12.2 Conditional analysis on all possible CYP2D6 diplotypes

Interestingly, when the endoxifen analysis is conditioned on the allele dosages of all possible *CYP2D6* diplotypes, the index SNP that remains is rs2413670 (*Figure 37*). This analysis tells us that when all other genetic factors of *CYP2D6* is accounted for, gene expression is the next largest genetic factor that influences endoxifen metabolism.

Figure 37: Regional association plot of endoxifen conditioned on all possible *CYP2D6* diplotypes

Regional association plots of endoxifen following conditional analysis on allele dosage of all possible *CYP2D6* diplotypes. Each point represents the p-value for the association between a genetic variant and gene expression. Index SNP for this analysis is rs2413670, which is the candidate proxy SNP selected for *CYP2D6* eQTL. Green point is rs5751245.

3.13 Endoxifen association testing conditional on rs2413670 allele dosage

Conditional analysis of *CYP2D6* eQTL on rs2413670 allele dosage revealed that it is an appropriate proxy for *CYP2D6* eQTLs because the association of rs28371699 was severely attenuated in *CYP2D6* and the association peak for *NDUFA6* disappeared altogether ([Figure 38](#)).

In the conditional analysis of endoxifen on rs2413670 allele dosage, some SNPs that were previously genome-wide significant such as the sentinel rs5751245 became severely attenuated, but still remained genome-wide significant ([Figure 39A](#) and [Figure 39B](#)). This phenomenon tells us that the genetic effects of rs2413670 is not separate from the effects of rs5751245.

Some SNPs that were previously not significantly associated with the endoxifen had overwhelmingly significant p-values after conditional analysis ([Figure 39C](#) and [Figure 39D](#)). It is highly unlikely that these SNPs are true associations as they are located around the same genomic location as the SNPs that were genome-wide significant in the original analysis, suggesting that this is an artefact of a smaller sample size.

Figure 38: Regional association plots of *CYP2D6* and *NDUFA6* eQTL Conditional Analysis

Regional association plots of *CYP2D6* and *NDUFA6* eQTL following conditional analysis on allele dosage on candidate proxy SNP rs2413670. Each point represents the p-value for the association between a genetic variant and gene expression. **A:** *CYP2D6* eQTL conditioned on rs2413670 allele dosage; **B:** *NDUFA6* eQTL conditioned on rs2413670 allele dosage.

Figure 39: Regional association plots of endoxifen conditional analysis on rs2413670

Before conditional analysis (**A**), SNPs highlighted in red were genome-wide significant and remained genome-wide significant after conditional analysis (**B**). Before conditional analysis (**C**) SNPs highlighted in yellow were not genome-wide significant but were genome-wide significant after conditional analysis (**D**).

3.14 SNP location and Open Chromatin Marks in Liver Tissue

To find overlaps between the genome-wide significant endoxifen SNPs, *CYP2D6* eQTL SNPs, and open chromatin regions, I analysed chromatin immunoprecipitation sequencing (ChIP-seq) datasets on the tripartite of open chromatin and active transcription histone marks; H3K27ac (acetylation of the 27th amino acid residue, lysine [K], of histone H3), H3K4me3 (trimethylation of the 4th residue K of histone H3), and H3K4me1 (monomethylation of the 4th residue K of histone H3).

We analysed a public dataset of the abovementioned three chromatin marks sequenced from healthy human liver tissue from GSE155908. Information on sequencing run, sample identification and accession numbers are detailed in [Table 15](#). Coverage tracks as well as the location of the genome-wide significant SNPs for endoxifen and *CYP2D6* eQTL were visualised using the Integrative Genomics Viewer (IGV), version (version 2.8.12).

At the location of the *CYP2D6* eQTL sentinel SNP rs28371699, we see an enrichment in H3K27ac as well as H3K4me3 marks. In the regions flanking the SNP are enrichment peaks for H3K4me1 ([Figure 40](#)). Together, the evidence strongly suggests that *CYP2D6* is a gene that harbours the hallmarks of an actively transcribed gene. Causality for rs28371699 has to be proven through in-vitro experiments, which is outside the scope of this thesis.

In contrast, the endoxifen sentinel SNP rs5751245 is in a genomic region that is absent of open chromatin marks in the liver tissue ([Figure 41](#)), suggesting that rs5751245 is indeed a proxy SNP that is in strong linkage disequilibrium with a yet-to-be-determined causal variant for varying endoxifen levels.

The proxy index SNP for *CYP2D6* eQTL rs2413670 is also a variant that is in a region where there is little enrichment of open chromatin histone modifications (*Figure 42*).

Figure 43 demonstrates rs5751245 as an upstream variant relative to *CYP2D6* transcription start site (TSS), and rs2413670 is located downstream relative to *CYP2D6* TSS.

Table 15: National Centre for Biotechnology Information Sequence Read Archive (NCBI-SRA) details for raw sequence files used from GSE155908.

Run	GEO_Accession	Tissue	Sample	Antibody
SRR12416642	GSM4715494	liver	N389	H3K27ac
SRR12416644	GSM4715496	liver	N389	H3K4me1
SRR12416645	GSM4715497	liver	N389	H3K4me3
SRR12416646	GSM4715498	liver	N389	input
SRR12416647	GSM4715499	liver	N396	H3K27ac
SRR12416649	GSM4715501	liver	N396	H3K4me1
SRR12416650	GSM4715502	liver	N396	H3K4me3
SRR12416651	GSM4715503	liver	N396	input
SRR12416652	GSM4715504	liver	N399	H3K27ac
SRR12416654	GSM4715506	liver	N399	H3K4me1
SRR12416655	GSM4715507	liver	N399	H3K4me3
SRR12416656	GSM4715508	liver	N399	H3input
SRR12416657	GSM4715509	liver	N400	H3K27ac
SRR12416659	GSM4715511	liver	N400	H3K4me1
SRR12416660	GSM4715512	liver	N400	input

Figure 40: IGV browser view of rs28371699

The blue line marks the genomic location of rs28371699. It lies in the genomic location with highly enriched peak signals of both H3K27ac (blue peaks) and H3K4me3 (red peaks). These chromatin marks indicate genomic areas of active transcription.

Figure 41: IGV browser view rs5751245

rs5751245 is a variant that does not lie in any open chromatin marks in liver tissue.

Figure 42: IGV browser view of rs2413670 and rs28371699

A: The red line marks the genomic location of rs2413670; **B:** the blue line marks the genomic location of rs28371699.

Figure 43: IGV browser view of rs2413670, rs28371699, and rs5751245

A: The red line marks the genomic location of rs2413670; **B:** the blue line marks the genomic location of rs28371699; **C:** the black line marks the genomic location of rs5751245

3.15 Linkage Disequilibrium between endoxifen, *CYP2D6* eQTL, and *CYP2D6* allele proxy SNPs

Statistically significant SNPs that are either not in linkage disequilibrium or have low degrees of linkage disequilibrium are more valuable when used in tandem in a linear regression. Hence, we determined the extent of linkage disequilibrium by calculating the pairwise R^2 between the endoxifen proxy SNP rs5751245, the *CYP2D6* eQTL proxy SNP rs2413670, and each of the most common *CYP2D6* alleles. The results are summarized in [Figure 44](#).

As expected, rs2413670 and rs5751245 is in moderately low linkage disequilibrium (Pairwise $R^2 = 0.28-0.34$) in all three endoxifen cohorts, contributing to the conditional analysis results that we observed in [Section 3.7.2](#). Also within expectation is the degree of linkage disequilibrium between the *CYP2D6* alleles and rs5751245 and rs2413670.

These results indicate two things: firstly, there is value in combining the allele dosages of rs5751245 and rs2413670 together into a linear regression analysis, and it will explain a greater degree of the variance than a single SNP alone. Secondly, rs5751245 and rs2413670 both explain a degree of the variance that is observed as a consequence of the *CYP2D6* alleles, as demonstrated by the conditional analyses ([Sections 3.7.2 and 3.13](#) respectively). The variance which the *CYP2D6* alleles account for is not equivalent to those explained by rs5751245 or rs2413670, but as demonstrated earlier in [Section 3.7.1](#) the amount of the variance explained is comparable.

SGP	rs2413670_C	rs1065852_A	rs5751245_C		Color Scale
rs2413670_C (0.69)	1				0
rs1065852_A (0.57)	0.45	1			0.1
rs5751245_C (0.41)	0.31	0.32	1		0.2
					0.3
DEU	rs2413670_C	rs3892097_T	rs5751245_C		0.4
rs2413670_C (0.54)	1				0.5
rs3892097_T (0.19)	0.2	1			0.6
rs5751245_C (0.29)	0.34	0.53	1		0.7
					0.8
LBN	rs2413670_C	rs28371725_T	rs5751245_C		0.9
rs2413670_C (0.54)	1				1
rs28371725_T (0.16)	0.16	1			
rs5751245_C (0.23)	0.28	0.14	1		

Figure 44: Pairwise R^2 between proxy SNPs and *CYP2D6* alleles

Row names indicate the following information: dbSNPID_allele (allele frequency). Value in cells indicates value of pairwise R^2 between two SNPs and the cell colour intensity follows the colour scale on the right, indicating the increasing intensity of red with increasing R^2 values. Zero indicates no correlation between two SNPs. 1 indicates perfect correlation and two SNPs are in complete linkage disequilibrium.

3.16 Improved predictive power using endoxifen and *CYP2D6* eQTL index SNPs

As demonstrated in [Section 3.7.1](#), rs5751245 is comparable in predictive power as compared to the most common *CYP2D6* alleles. Similarly, rs2413670 also explains approximately 18-24% of the variance observed in serum endoxifen levels. Including a second variant into the linear regression equation increases the predictive power of the regression model substantially, explaining between 24-30% of the variance observed. We tested the combinations of rs5751245 with rs2413670, rs5751245 with each cohort's most common *CYP2D6* allele, and finally, the combination of all three variants. The results are tabulated in [Table 16](#)**Error! Reference source not found.** The multi-variant regression tests assumes that each genetic variant carries the same weight, ergo each copy of an alternative allele carries the same weight.

In DEU, rs5751245 and *CYP2D6**4 together explains approximately 24% of the variance observed. rs5751245 in combination with rs2413670 performed marginally better, explaining 26% of the variance. The three-variant combination demonstrated a marginal increase in the amount of variance explained, increasing to 27%. In SGP, rs5751245 with *CYP2D6**10 explained approximately 25% of the variance, whereas rs5751245 with rs2413670 explained a comparable 24%. SGP's three variant combination explained the most variance, at 26%. In LBN, rs5751245 and *CYP2D6**41 together explained 26% of the variance, rs5751245 and rs2413670 explained 29% of the variance, whereas the three variant combination surprisingly performed intermediate of the two-variant combinations, at 28% of the variance explained. The statistical significance

between the two- and three-variant combinations are highly similar within each cohort ([Table 17](#)). The results of the multi-variant regression analysis demonstrate that the three-variant regression model does not give a substantial increase in predictive power over the two-variant model; only 1-2% increase in variance explained.

Considering the similar performance of the two two-variant regression models, the one with rs5751245 and rs2413670 is preferable considering the fact that rs2413670 and rs5751245 are genetic variants that are common in all tested ethnic populations, versus the drastically different allele frequencies of the *CYP2D6* alleles ([Table 16](#), [Table 1](#) and [Table 2](#)).

In the final meta-analysis, the two-variant model consisting of rs5751245 and rs2413670 was overwhelmingly significant ($p_{2SNP} = 1.1 \times 10^{-34}$) and more statistically significant than either of the two SNPs individually ([Table 17](#)). The counts and the frequencies of the SNP combinations for rs2413670 and rs5751245 found in our study are summarised in [Table 18](#).

When comparing the two-SNP model to the diplotype-derived AS model as described in [Section 3.5](#), the AS model is of greater statistical significance ($p_{AS} = 3.65 \times 10^{-45}$) ([Table 9](#)). In SGP, the AS model explains 23% of the variance whereas the two-SNP model explains 24% of the variance. In DEU, the AS model performs better, explaining 31% of the variance compared to 26% of the two-SNP model. In LBN, the two-SNP model performs better, explaining 30% of the variance compared to 26% of the AS model. The results demonstrate that the two-SNP model's performance is comparable to the AS model.

Table 16: Linear regression results of rs5751245 and rs2413670 allele dosage on endoxifen and meta-analysis

	β	SE	T	P	R^2	Allele Freq	β_{meta}	SE _{meta}	P _{meta}
CYP2D6 alleles									
SGP <i>CYP2D6</i> *10 rs1065852_A	-0.1485	0.02432	-6.107	8.73E-09	0.21	0.57	-	-	-
DEU <i>CYP2D6</i> *4 rs3892097_T	-0.1961	0.0251	-7.814	1.10E-13	0.18	0.19			
LBN <i>CYP2D6</i> *4I rs28371725_T	-0.0889	0.06183	-1.439	0.16	0.17	0.16			
rs5751245_C									
SGP	-0.1618	0.028	-5.776	4.40E-08	0.18	0.41	-0.174	0.016	1.27E-28
DEU	-0.1795	0.0202	-8.876	8.47E-17	0.22	0.29			
LBN	-0.1794	0.0533	-3.367	1.39E-03	0.18	0.23			
rs2413670_C									
SGP	-0.1552	0.0287	-5.409	2.53E-07	0.18	0.69	-0.147	0.015	1.21E-23
DEU	-0.152	0.0189	-8.038	2.50E-14	0.19	0.54			
LBN	-0.111	0.0405	-2.739	8.21E-03	0.24	0.54			
rs5751245_C + rs2413670_C									
SGP	-0.1059	0.01576	-6.719	3.81E-10	0.24	-	-0.102	0.008	1.14E-34
DEU	-0.1024	0.01052	-9.73	1.66E-19	0.26	-			
LBN	-0.0872	0.02596	-3.359	1.42E-03	0.30	-			
<i>R² refers to amount of variance the regression model explains. Bolded row names indicate dbSNPID allele</i>									

Table 17: Linear Regression summary statistics of rs5751245, the most common *CYP2D6* allele in each respective cohort, and rs2413670 on endoxifen

	β	SE	T	P	R^2
DEU					
rs5751245 + <i>CYP2D6</i> *4	-0.109	0.01178	-9.25	5.46E-18	0.24
rs5751245 + rs2413670	-0.1024	0.01052	-9.73	1.66E-19	0.26
rs5751245 + <i>CYP2D6</i> *4 + rs2413670	-0.0787	0.00781	-10.07	1.31E-20	0.27
Activity Score	0.2352	0.0206	11.4325	3.31E-25	0.31
SGP					
rs5751245 + <i>CYP2D6</i> *10	-0.0958	0.01396	-6.98	1.75E-10	0.25
rs5751245 + rs2413670	-0.1059	0.01576	-6.72	3.81E-10	0.24
rs5751245 + <i>CYP2D6</i> *10 + rs2413670	-0.0694	0.00994	-6.99	9.02E-11	0.26
Activity Score	0.2051	0.0305	6.7252	3.57E-10	0.23
LBN					
rs5751245 + <i>CYP2D6</i> *41	-0.066	0.02031	-3.25	1.93E-03	0.26
rs5751245 + rs2413670	-0.0872	0.02596	-3.36	1.43E-03	0.30
rs5751245 + <i>CYP2D6</i> *41 + rs2413670	-0.066	0.02031	-3.25	1.93E-03	0.28
Activity Score	0.2313	0.0476	4.8605	7.39E-06	0.26
<i>R² refers amount of the variance explained by the regression model.</i>					

Table 18: Alternate Allele Counts of rs2413670_C and rs5751245_C and frequency

Allele Count	SGP		DEU		LBN	
rs2413670_C/rs5751245_C (Total Count)	Count	Freq	Count	Freq	Count	Freq
0/0 (0)	20	0.13	67	0.23	18	0.26
0/1 (1)	0	0.00	0	0.00	0	0.00
0/2 (2)	0	0.00	0	0.00	0	0.00
1/0 (1)	15	0.10	64	0.22	16	0.23
1/1 (2)	38	0.25	67	0.23	13	0.19
1/2 (3)	0	0.00	1	0.00	0	0.00
2/0 (2)	18	0.12	18	0.06	5	0.07
2/1 (3)	34	0.22	41	0.14	12	0.17
2/2 (4)	26	0.17	29	0.10	3	0.04
Missing	1	0.01	3	0.01	2	0.03

0 allele count indicates reference allele, 1 or 2 indicates number of copies of the indicated alternate allele.

3.17 Other *CYP2D6* alleles in other studies

In our study, we have managed to either directly genotype, or impute with high stringency some SNPs that have been found to have association with either drug metabolism or with *CYP2D6* expression.

Yang et al.¹⁵⁹ tested for and found associations in three SNPs ($p < 1e-11$) with both differential *CYP2D6* expression and dextromethorphan metabolism in Caucasian cohorts: rs8138080, rs5751247, and rs17478227. In our study, all three SNPs were found to be associated with endoxifen ($p < 1e-16$), but association with *CYP2D6* eQTL was only nominal (**Error! Reference source not found.**).

rs8138080 was not associated with endoxifen in the SGP cohort ($p = 0.6$) whereas strong association was found in our DEU cohort ($p = 8.95 \times 10^{-21}$). Its association with *CYP2D6* eQTL in the SCRB cohort was also only nominal ($p = 0.02$). rs8138080 was also the sentinel SNP in Hennig et al.'s¹⁵⁷ study.

rs5751247 is strongly associated with endoxifen, but it did not pass quality control measures in the SCRB cohort and hence not tested for association with eQTLs. However, rs5751247 is in tight LD with our endoxifen proxy SNP rs5751245 (pairwise $R^2 = 0.93$) according to LDLink²¹³. rs5751245 was associated with *CYP2D6* eQTL, but it was only moderately significant ($p = 5 \times 10^{-4}$). It stands to reason that rs5751247 is also not strongly associated with *CYP2D6* eQTL in our study either.

Similarly, rs17478227 did not pass quality control measures in SGP for endoxifen. Its allele frequency is also rare in the SCRB cohort ($MAF=0.02$) and no association with *CYP2D6* eQTL was found ($p = 0.3$). Indeed, rs8138080 and

rs17478227 are both rare and are not good proxy SNPs for the East Asian (EAS) population ([Figure 45](#)**Error! Reference source not found.**).

Wang et al.²¹⁴ described rs5758550 marking an enhancer region that is able to almost double the amount of *CYP2D6* transcription in HepG2 cells and suggests, its inclusion into the array of *CYP2D6* alleles to increase predictive resolution. rs5758550 was found to be in high degree of linkage disequilibrium with *CYP2D6**2 allele. On the contrary, Sanchez-Spitman et al.²¹⁵ claims that rs5758550 does not increase the amount of variance explained in endoxifen. In our study, rs5758550 is neither associated with endoxifen ($p = 0.011$) nor *CYP2D6* eQTL ($p = 0.7$) (**Error! Reference source not found.**). Although we cannot discount the possibility that rs5758550 is ethnic-specific and it has no effect on *CYP2D6* expression in an Asian cohort, the evidence is very strong that it has no effect on endoxifen metabolism - rs5758550 is very likely a false-positive.

Figure 45: Pairwise R^2 of SNPs from Yang et al. (2010)¹⁵⁹ and *CYP2D6*4* and *CYP2D6*10* in CEU (Caucasian) and EAS (East Asian) populations from LDLink Database²¹³

Column and Row headers contains the dbSNPID(MAF). *CYP2D6*4* proxy rs3892097; *CYP2D6*10* proxy rs1065852. Value in cells indicates value of pairwise R^2 between two SNPs. Colour scale on the right indicates the increasing intensity of the red colouring with increasing R^2 values. 0 indicates no correlation between two SNPs and 1 indicates perfect correlation and two SNPs are in complete linkage disequilibrium.

Table 19: SNPs found in other studies

Chr	BP	SNP	Alt Allele	Cohort	Endoxifen Assoc			Endo Meta			SCRB eQTL		
					MAF	β	P	β	P	I ²	MAF	β	P
22	42396371	rs8138080 ^{157,159}	A	DEU	0.29	-0.1964	8.90E-21	-0.1835	7.90E-27	65.09	0.1	-2.194	0.02732
				LBN	0.3	-0.1843	8.50E-05						
				SGP	0.04	-0.03062	0.6458						
22	42412711	rs5758550 ^{159,214,215}	G	DEU	0.2232	0.04	0.1128	0.05	0.01	0	0.09	0.13	0.7
				LBN	0.2464	0.06	0.2442						
				SGP	0.0724	0.1	0.0562						
22	42633054	rs5751247 ¹⁵⁹	C	DEU	0.28	-0.1835	2.06E-17	-0.1716	8.99E-28	0	-	-	-
				LBN	0.22	-0.1728	0.001						
				SGP	0.4	-0.1478	6.01E-07						
22	42654327	rs17478227 ¹⁵⁹	G	DEU	0.18	-0.1962	1.60E-12	-0.2031	1.55E-16	0	0.02	-2.203	0.3221
				LBN	0.1	-0.2471	0.0005						
				SGP	-	-	-						

4 Discussion

Our study is the first multi-ethnic GWAS that attempts to elucidate the precise genetic underpinnings of tamoxifen's metabolism to endoxifen, by combining results from testing for association with endoxifen as well as expression quantitative trait loci.

4.1 Genetic variants other than *CYP2D6* are not associated with steady state endoxifen

Tamoxifen's metabolism is highly complex and involves several Phase I and II drug metabolizing enzymes. Endoxifen is a secondary metabolite derived from two pathways; the major pathway is the 4-hydroxylation of NDM by CYP2D6, the secondary pathway is the *N*-desmethylation of 4OHT by CYP3A4/5⁴⁶. Other cytochrome P450s CYP2C9, CYP2C19, CYP1A2, and CYP2B6 have been demonstrated to have metabolic activity as well, but their effects have been residual⁵¹. Phase II enzymes such as UGTs also affects the metabolism of tamoxifen overall. In this multi-ethnic GWAS analysis, these candidate loci did not demonstrate even suggestive significance ($p < 1e-5$). This tells us that common genetic variations at these loci have minimal contribution to serum endoxifen levels.

4.2 GWAS reveals rs5751245 as sentinel SNP for steady state endoxifen

I have identified rs5751245 as a genetic variant strongly associated with lower average amounts of steady state endoxifen levels, indicating a clear gene-

exposure effect. rs5751245 demonstrated no evidence of heterogeneity between the multi-ethnic cohorts of Singapore, Germany, and Lebanon. Evidence from ChIP-seq of open chromatin histone marks further support that rs5751245 is merely an excellent proxy variant that is in linkage disequilibrium with haplotypes of *CYP2D6* with the largest effects, namely *CYP2D6**4, *CYP2D6**10, and *CYP2D6**41. However, rs5751245 is superior in its utility to the *CYP2D6* alleles as it is a common variant in all the ethnic cohorts tested and carries as much information as any of the individual *CYP2D6* alleles and can be genotyped in all patients regardless of ethnicity.

I have also identified rs2413670 as a proxy variant for mediating *CYP2D6* gene expression; our eQTL GWAS demonstrating a clear gene expression quantitative trait loci. Furthermore, this variant is also strongly associated with steady state serum endoxifen levels that demonstrates no heterogeneity amongst the ethnic cohorts that I tested. It is also a common variant that is polymorphic across all studied patient cohorts. The results of the association analysis conditional on rs2413670 allele dosage reveals that it is only responsible for a portion of the overall genetic effect exerted on steady state endoxifen levels.

I have demonstrated an additive effect of these two genetic variants on steady state serum endoxifen levels. Together, rs5751245 and rs2413670 explain more than 26% of the variance observed in endoxifen levels. Comparing the rs5751245 and rs2413670 two-SNP model to the *CYP2D6* activity score derived linear regression model, the two different models are highly similar in terms of explaining the amount of variability in steady state serum endoxifen levels. The two-SNP model is more practical as the SNPs involved are common variants in

all ethnic populations, and requires only genotyping two SNPs, reducing the cost of genotyping by half over the AS model, where at least 4 variants are required.

Our study is in agreement with the findings of Yang et al.¹⁵⁹ with regard to metabolism of a drug known to be mediated by CYP2D6. Our expression analysis was conducted in a different ethnic group (Asian vs Caucasian). Hence, it is within expectation that differences in allele frequencies had led to differential findings, as demonstrated in [Section 3.17](#). Taken together, the evidence strongly suggests the presence of ethnic specific mechanisms governing *CYP2D6* expression, and indeed there are proxy SNPs that are strongly associated with CYP2D6 drug metabolism.

4.3 rs5751245 annotated as an *TCF20* upstream variant

Remarkably, our sentinel SNP rs5751245 is a distal genetic variant some 74 kb upstream of the *CYP2D6* TSS, ~9.4 kb upstream of the *TCF20* TSS. *TCF20* was never implicated in tamoxifen metabolism. Previously, it has been identified as a co-regulator of other transcription factors such as c-Jun, Sp1 (Specificity protein 1), Ets1 (Protein C-ets-1), Pax6 (Paired box protein 6)²¹⁶, RNF4 (Ring finger protein 4)²¹⁷, AR (androgen receptor)²¹⁸, and the ER α ²¹⁹. Gburcik et al.²¹⁹ demonstrated that *TCF20* interacts specifically with activated ER α in the ER+ MCF7 cell line and overexpression of *TCF20* resulted in a significant reduction of colonies. They further demonstrated that *TCF20* works specifically with ER α because in the SkBr3 cell line (an ER-negative breast cancer cell line), overexpression of *TCF20* did not affect colony formation. There are no further studies conducted regarding *TCF20* and ER α . The possibility that *TCF20* explains further the variability of steady state endoxifen is remote at best.

4.4 Conclusion

Our results demonstrate that there are genetic variants that are strongly associated with serum endoxifen levels and *CYP2D6* eQTL. *CYP2D6* expression explains a portion of the variance observed in serum endoxifen levels but is still a secondary factor. The strongest association with serum endoxifen levels, rs5751245, is located a significant distance (~74 kb) away from *CYP2D6*, in a genomic location that lacks open chromatin marks in liver tissue. Collectively, the evidence suggests that rs5751245 is serving as a proxy, in linkage disequilibrium with the causal variant or haplotype. Combining the genetic variants for endoxifen and *CYP2D6* eQTL in a linear regression model explains more than 26% of the variance observed in steady state serum endoxifen levels in ethnic diverse patient cohorts. The performance of the linear regression model derived from our GWAS results are comparable to the current activity score derived system.

4.5 Suggestions for future work

Further investigation into the genomic architecture of the region can be carried out to pinpoint the causal variant to endoxifen's variation. Sequencing with long reads such as Pacific Biosciences' single molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies' nanopore sequencing has allowed the identification of new haplotypes, identification of structural variants, as well as refinements in variant calling to determine *CYP2D6* diplotype^{220–222}. Chromatin accessibility techniques such as Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) can be employed to further refine the regional map that governs *CYP2D6* gene expression. In addition, an enzyme-

linked immunosorbent assay (ELISA) or mass spectrometry can also be employed to measure amounts of CYP2D6 protein present and to test for association with genetic variants.

Without the recruitment of larger sample sizes, it is difficult to conduct association testing for rare variants, or for the detection of *de novo* variants. In a perfect world, it will be most optimal to conduct multiple time point tamoxifen metabolism measurement, liver tissue eQTL, as well as targeted sequencing of the *CYP2D6* loci all in one large study cohort. It is with great hope to be able to replicate the results in a separate cohort.

5 Bibliography

1. Weinshilboum, R. Inheritance and Drug Response. *N. Engl. J. Med.* **348**, 529–537 (2003).
2. Evans, W. E. & McLeod, H. L. Pharmacogenomics — Drug Disposition, Drug Targets, and Side Effects. *N. Engl. J. Med.* **348**, 538–549 (2003).
3. Mahgoub, A., Dring, L. G., Idle, J. R., Lancaster, R. & Smith, R. L. Polymorphic hydroxylation of debrisoquine in man. *Orig. Publ. Vol. 2 Issue 8038* **310**, 584–586 (1977).
4. Eichelbaum, M., Spannbrucker, N., Steincke, B. & Dengler, H. J. Defective N-oxidation of sparteine in man: A new pharmacogenetic defect. *Eur. J. Clin. Pharmacol.* **16**, 183–187 (1979).
5. Bertilsson, L. *et al.* Nortriptyline and antipyrine clearance in relation to debrisoquine hydroxylation in man. *Life Sci.* **27**, 1673–1677 (1980).
6. Schmid, B., Bircher, J., Preisig, R. & Kpfer, A. Polymorphic dextromethorphan metabolism: Co-segregation of oxidative O-demethylation with debrisoquin hydroxylation. *Clin. Pharmacol. Ther.* **38**, 618–624 (1985).
7. Dahl-Puustinen, M. L., Lidn, A., Alm, C., Nordin, C. & Bertilsson, L. Disposition of perphenazine is related to polymorphic debrisoquin hydroxylation in human beings. *Clin. Pharmacol. Ther.* **46**, 78–81 (1989).
8. Distlerath, L. M. & Guengerich, F. P. Characterization of a human liver cytochrome P-450 involved in the oxidation of debrisoquine and other drugs by using antibodies raised to the analogous rat enzyme. *Proc. Natl. Acad. Sci.* **81**, 7348–7352 (1984).

9. Eichelbaum, M. *et al.* Chromosomal assignment of human cytochrome P-450 (debrisoquine/sparteine type) to chromosome 22. *Br. J. Clin. Pharmacol.* **23**, 455–458 (1987).
10. Kimura, S., Umeno, M., Skoda, R. C., Meyer, U. A. & Gonzalez, F. J. The human debrisoquine 4-hydroxylase (CYP2D) locus: sequence and identification of the polymorphic CYP2D6 gene, a related gene, and a pseudogene. *Am. J. Hum. Genet.* **45**, 889–904 (1989).
11. Heim, M. & Meyer, U. A. Genotyping of poor metabolisers of debrisoquine by allele-specific PCR amplification. *The Lancet* **336**, 529–532 (1990).
12. Daly, A. K. *et al.* Nomenclature for human CYP2D6 alleles. *Pharmacogenet. Genomics* **6**, 193–201 (1996).
13. Ingelman-Sundberg, M., Oscarson, M., Daly, A. K., Garte, S. & Nebert, D. W. Human Cytochrome P-450 (CYP) Genes: A Web Page for the Nomenclature of Alleles. *Cancer Epidemiol. Prev. Biomark.* **10**, 1307–1308 (2001).
14. Gaedigk, A. *et al.* The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. *Clin. Pharmacol. Ther.* **103**, 399–401 (2018).
15. Whirl-Carrillo, M. *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **92**, 414–417 (2012).
16. Goetz, M. P. *et al.* Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for CYP2D6 and Tamoxifen Therapy. *Clin. Pharmacol. Ther.* **103**, 770–777 (2018).

17. Flockhart, D. A. Drug Interactions: Cytochrome P450 Drug Interaction Table. *Indiana University School of Medicine* <https://drug-interactions.medicine.iu.edu> (2007).
18. Harper, M. J. & Walpole, A. L. A new derivative of triphenylethylene: effect on implantation and mode of action in rats. *J. Reprod. Fertil.* **13**, 101–119 (1967).
19. Lunan, C. B. & Klopper, A. Antioestrogens: A Review. *Clin. Endocrinol. (Oxf.)* **4**, 551–572 (1975).
20. Cole, M. P., Jones, C. T. A. & Todd, I. D. H. A New Anti-oestrogenic Agent in Late Breast Cancer: An Early Clinical Appraisal of ICI46474. *Br. J. Cancer* **25**, 270–275 (1971).
21. Ward, H. W. C. Anti-oestrogen Therapy for Breast Cancer: A Trial of Tamoxifen at Two Dose Levels. *Br Med J* **1**, 13–14 (1973).
22. Jensen, E. V., Block, G. E., Smith, S., Kyser, K. & DeSombre, E. R. Estrogen receptors and breast cancer response to adrenalectomy. *Natl. Cancer Inst. Monogr.* **34**, 55–70 (1971).
23. McGuire, W. L., Carbone, P. P. & Vollmer, E. P. *Estrogen receptors in human breast cancer*. (Raven Press, 1975).
24. JORDAN, V. ANTITUMOR ACTIVITY OF ANTIESTROGEN ICI 46,474 (TAMOXIFEN) IN DIMETHYLBENZANTHRACENE (DMBA)-INDUCED RAT MAMMARY-CARCINOMA MODEL. *J. Steroid Biochem. Mol. Biol.* **5**, 354–354 (1974).

25. JORDAN, V. & KOERNER, S. TAMOXIFEN (ICI 46,474) AND HUMAN CARCINOMA 8S ESTROGEN-RECEPTOR. *Eur. J. Cancer* **11**, 205–206 (1975).
26. Jordan, V. C. Effect of tamoxifen (ICI 46,474) on initiation and growth of DMBA-induced rat mammary carcinomata. *Eur. J. Cancer* **12**, 419–424 (1976).
27. Jordan, V. C. & Dowse, L. J. TAMOXIFEN AS AN ANTI-TUMOUR AGENT EFFECT ON OESTROGEN BINDING. *J. Endocrinol.* **68**, 297–303 (1976).
28. Jordan, V. C. Tamoxifen as the first targeted long-term adjuvant therapy for breast cancer. *Endocr. Relat. Cancer* **21**, R235–R246 (2014).
29. Early Breast Cancer Trialists' Collaborative Group. Effects of Adjuvant Tamoxifen and of Cytotoxic Therapy on Mortality in Early Breast Cancer. *N. Engl. J. Med.* **319**, 1681–1692 (1988).
30. Early Breast Cancer Trialist's Collaborative Group. Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy: 133 randomised trials involving 31 000 recurrences and 24 000 deaths among 75 000 women. *The Lancet* **339**, 71–85 (1992).
31. Early Breast Cancer Trialist's Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials. *The Lancet* **351**, 1451–1467 (1998).
32. Early Breast Cancer Trialist's Collaborative Group. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and

- 15-year survival: an overview of the randomised trials. *The Lancet* **365**, 1687–1717 (2005).
33. Davies, C. *et al.* Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: ATLAS, a randomised trial. *The Lancet* **381**, 805–816 (2013).
 34. Riggs, B. L. & Hartmann, L. C. Selective Estrogen-Receptor Modulators — Mechanisms of Action and Application to Clinical Practice. *N. Engl. J. Med.* **348**, 618–629 (2003).
 35. Clemons, M. & Goss, P. Estrogen and the Risk of Breast Cancer. *N. Engl. J. Med.* **344**, 276–285 (2001).
 36. Osborne, C. K. Tamoxifen in the Treatment of Breast Cancer. *N. Engl. J. Med.* **339**, 1609–1618 (1998).
 37. Dahlman-Wright, K. *et al.* International Union of Pharmacology. LXIV. Estrogen receptors. *Pharmacol. Rev.* **58**, 773–781 (2006).
 38. Heldring, N. *et al.* Estrogen receptors: how do they signal and what are their targets. *Physiol. Rev.* **87**, 905–931 (2007).
 39. Carlson, R. W. The History and Mechanism of Action of Fulvestrant. *Clin. Breast Cancer* **6**, S5–S8 (2005).
 40. Wijayaratne, A. L. & McDonnell, D. P. The Human Estrogen Receptor- α Is a Ubiquitinated Protein Whose Stability Is Affected Differentially by Agonists, Antagonists, and Selective Estrogen Receptor Modulators. *J. Biol. Chem.* **276**, 35684–35692 (2001).

41. Maximov, P. Y. *et al.* Endoxifen, 4-Hydroxytamoxifen and an Estrogenic Derivative Modulate Estrogen Receptor Complex Mediated Apoptosis in Breast Cancer. *Mol. Pharmacol.* **94**, 812–822 (2018).
42. Jameera Begam, A., Jubie, S. & Nanjan, M. J. Estrogen receptor agonists/antagonists in breast cancer therapy: A critical review. *Bioorganic Chem.* **71**, 257–274 (2017).
43. Lonard, D. M., Nawaz, Z., Smith, C. L. & O'Malley, B. W. The 26S Proteasome Is Required for Estrogen Receptor- α and Coactivator Turnover and for Efficient Estrogen Receptor- α Transactivation. *Mol. Cell* **5**, 939–948 (2000).
44. Lim, Y. C., Desta, Z., Flockhart, D. A. & Skaar, T. C. Endoxifen (4-hydroxy-N-desmethyl-tamoxifen) has anti-estrogenic effects in breast cancer cells with potency similar to 4-hydroxy-tamoxifen. *Cancer Chemother. Pharmacol.* **55**, 471–478 (2005).
45. Lim, Y. C. *et al.* Endoxifen, a secondary metabolite of tamoxifen, and 4-OH-tamoxifen induce similar changes in global gene expression patterns in MCF-7 breast cancer cells. *J. Pharmacol. Exp. Ther.* **318**, 503–512 (2006).
46. 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 (2004).
47. Wu, X. *et al.* The Tamoxifen Metabolite, Endoxifen, Is a Potent Antiestrogen that Targets Estrogen Receptor α for Degradation in Breast Cancer Cells. *Cancer Res.* **69**, 1722–1727 (2009).

48. Hawse, J. R. *et al.* Endoxifen's Molecular Mechanisms of Action Are Concentration Dependent and Different than That of Other Anti-Estrogens. *PLOS ONE* **8**, e54613 (2013).
49. Klein, D. J. *et al.* PharmGKB summary: tamoxifen pathway, pharmacokinetics. *Pharmacogenet. Genomics* **23**, 643–647 (2013).
50. Dehal, S. S. & Kupfer, D. CYP2D6 Catalyzes Tamoxifen 4-Hydroxylation in Human Liver. *Cancer Res.* **57**, 3402–3406 (1997).
51. 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).
52. Stearns, V. *et al.* Active tamoxifen metabolite plasma concentrations after coadministration of tamoxifen and the selective serotonin reuptake inhibitor paroxetine. *J. Natl. Cancer Inst.* **95**, 1758–1764 (2003).
53. Jin, Y. *et al.* CYP2D6 Genotype, Antidepressant Use, and Tamoxifen Metabolism During Adjuvant Breast Cancer Treatment. *JNCI J. Natl. Cancer Inst.* **97**, 30–39 (2005).
54. Teft, W. A. *et al.* CYP3A4 and seasonal variation in vitamin D status in addition to CYP2D6 contribute to therapeutic endoxifen level during tamoxifen therapy. *Breast Cancer Res. Treat.* **139**, 95–105 (2013).
55. Borges, S. *et al.* Quantitative effect of CYP2D6 genotype and inhibitors on tamoxifen metabolism: implication for optimization of breast cancer treatment. *Clin. Pharmacol. Ther.* **80**, 61–74 (2006).

56. Madlensky, L. *et al.* Tamoxifen Metabolite Concentrations, CYP2D6 Genotype, and Breast Cancer Outcomes. *Clin. Pharmacol. Ther.* **89**, 718–725 (2011).
57. Saladores, P. *et al.* Tamoxifen metabolism predicts drug concentrations and outcome in premenopausal patients with early breast cancer. *Pharmacogenomics J* **15**, 84–94 (2015).
58. Schroth, W. *et al.* Breast cancer treatment outcome with adjuvant tamoxifen relative to patient CYP2D6 and CYP2C19 genotypes. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **25**, 5187–5193 (2007).
59. Schroth, W. *et al.* Association between CYP2D6 polymorphisms and outcomes among women with early stage breast cancer treated with tamoxifen. *JAMA* **302**, 1429–1436 (2009).
60. Rae, J. M. *et al.* CYP2D6 and UGT2B7 genotype and risk of recurrence in tamoxifen-treated breast cancer patients. *J. Natl. Cancer Inst.* **104**, 452–460 (2012).
61. Regan, M. M. *et al.* CYP2D6 genotype and tamoxifen response in postmenopausal women with endocrine-responsive breast cancer: the breast international group 1-98 trial. *J. Natl. Cancer Inst.* **104**, 441–451 (2012).
62. de Vries Schultink, A. H. M., Zwart, W., Linn, S. C., Beijnen, J. H. & Huitema, A. D. R. Effects of Pharmacogenetics on the Pharmacokinetics and Pharmacodynamics of Tamoxifen. *Clin. Pharmacokinet.* **54**, 797–810 (2015).
63. Erratum. *JNCI J. Natl. Cancer Inst.* **104**, 1772–1772 (2012).

64. Goetz, M. P. *et al.* CYP2D6 metabolism and patient outcome in the Austrian Breast and Colorectal Cancer Study Group trial (ABCSG) 8. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **19**, 500–507 (2013).
65. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* **9**, 2941 (2018).
66. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).
67. Meyre, D. *et al.* Variants of ENPP1 are associated with childhood and adult obesity and increase the risk of glucose intolerance and type 2 diabetes. *Nat. Genet.* **37**, 863–867 (2005).
68. Meyre, D. *et al.* A Genome-Wide Scan for Childhood Obesity–Associated Traits in French Families Shows Significant Linkage on Chromosome 6q22.31–q23.2. *Diabetes* **53**, 803–811 (2004).
69. Maddux, B. A. *et al.* Membrane glycoprotein PC-1 and insulin resistance in non-insulin-dependent diabetes mellitus. *Nature* **373**, 448–451 (1995).
70. Weedon, M. N. *et al.* No evidence of association of ENPP1 variants with type 2 diabetes or obesity in a study of 8,089 U.K. Caucasians. *Diabetes* **55**, 3175–3179 (2006).
71. Lyon, H. N. *et al.* Common variants in the ENPP1 gene are not reproducibly associated with diabetes or obesity. *Diabetes* **55**, 3180–3184 (2006).
72. Grant, S. F. A. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320–323 (2006).

73. Reynisdottir, I. *et al.* Localization of a Susceptibility Gene for Type 2 Diabetes to Chromosome 5q34–q35.2. *Am. J. Hum. Genet.* **73**, 323–335 (2003).
74. Saxena, R. *et al.* Common single nucleotide polymorphisms in TCF7L2 are reproducibly associated with type 2 diabetes and reduce the insulin response to glucose in nondiabetic individuals. *Diabetes* **55**, 2890–2895 (2006).
75. Groves, C. J. *et al.* Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes* **55**, 2640–2644 (2006).
76. Scott, L. J. *et al.* Association of transcription factor 7-like 2 (TCF7L2) variants with type 2 diabetes in a Finnish sample. *Diabetes* **55**, 2649–2653 (2006).
77. van Vliet-Ostaptchouk, J. V. *et al.* Association of variants of transcription factor 7-like 2 (TCF7L2) with susceptibility to type 2 diabetes in the Dutch Breda cohort. *Diabetologia* **50**, 59–62 (2007).
78. Hayashi, T., Iwamoto, Y., Kaku, K., Hirose, H. & Maeda, S. Replication study for the association of TCF7L2 with susceptibility to type 2 diabetes in a Japanese population. *Diabetologia* **50**, 980–984 (2007).
79. Parra, E. J. *et al.* Association of TCF7L2 polymorphisms with type 2 diabetes in Mexico City. *Clin. Genet.* **71**, 359–366 (2007).
80. Bodhini, D., Radha, V., Dhar, M., Narayani, N. & Mohan, V. The rs12255372(G/T) and rs7903146(C/T) polymorphisms of the TCF7L2 gene

- are associated with type 2 diabetes mellitus in Asian Indians. *Metabolism*. **56**, 1174–1178 (2007).
81. Chang, Y.-C. *et al.* Association study of the genetic polymorphisms of the transcription factor 7-like 2 (TCF7L2) gene and type 2 diabetes in the Chinese population. *Diabetes* **56**, 2631–2637 (2007).
82. Ng, M. C. Y. *et al.* Replication and identification of novel variants at TCF7L2 associated with type 2 diabetes in Hong Kong Chinese. *J. Clin. Endocrinol. Metab.* **92**, 3733–3737 (2007).
83. Amoli, M. M. *et al.* Replication of TCF7L2 rs7903146 association with type 2 diabetes in an Iranian population. *Genet. Mol. Biol.* **33**, 449–451 (2010).
84. Alami, F. M. *et al.* Association of the TCF7L2 rs12255372 (G/T) variant with type 2 diabetes mellitus in an Iranian population. *Genet. Mol. Biol.* **35**, 413–417 (2012).
85. Schymick, J. C. *et al.* Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol.* **6**, 322–328 (2007).
86. Dunckley, T. *et al.* Whole-genome analysis of sporadic amyotrophic lateral sclerosis. *N. Engl. J. Med.* **357**, 775–788 (2007).
87. van Es, M. A. *et al.* ITPR2 as a susceptibility gene in sporadic amyotrophic lateral sclerosis: a genome-wide association study. *Lancet Neurol.* **6**, 869–877 (2007).
88. van Es, M. A. *et al.* Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.* **40**, 29–31 (2008).

89. Cronin, S. *et al.* A genome-wide association study of sporadic ALS in a homogenous Irish population. *Hum. Mol. Genet.* **17**, 768–774 (2008).
90. van Es, M. A. *et al.* Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nat. Genet.* **41**, 1083–1087 (2009).
91. Fogh, I. *et al.* No association of DPP6 with amyotrophic lateral sclerosis in an Italian population. *Neurobiol. Aging* **32**, 966–967 (2011).
92. Kwee, L. C. *et al.* A high-density genome-wide association screen of sporadic ALS in US veterans. *PloS One* **7**, e32768 (2012).
93. Shatunov, A. *et al.* Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: A genome-wide association study. *Lancet Neurol.* **9**, 986–994 (2010).
94. Laaksovirta, H. *et al.* Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study. *Lancet Neurol.* **9**, 978–985 (2010).
95. Gijssels, I. *et al.* Identification of 2 Loci at chromosomes 9 and 14 in a multiplex family with frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Arch. Neurol.* **67**, 606–616 (2010).
96. Renton, A. E. *et al.* A Hexanucleotide Repeat Expansion in C9ORF72 Is the Cause of Chromosome 9p21-Linked ALS-FTD. *Neuron* **72**, 257–268 (2011).
97. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron* **72**, 245–256 (2011).

98. Gijssels, I. *et al.* A C9orf72 promoter repeat expansion in a Flanders-Belgian cohort with disorders of the frontotemporal lobar degeneration-amyotrophic lateral sclerosis spectrum: a gene identification study. *Lancet Neurol.* **11**, 54–65 (2012).
99. Tsai, C.-P. *et al.* A hexanucleotide repeat expansion in C9ORF72 causes familial and sporadic ALS in Taiwan. *Neurobiol. Aging* **33**, 2232.e11–2232.e18 (2012).
100. Cintra, V. P. *et al.* The frequency of the C9orf72 expansion in a Brazilian population. *Neurobiol. Aging* **66**, 179.e1–179.e4 (2018).
101. Gromicho, M. *et al.* Frequency of C9orf72 hexanucleotide repeat expansion and SOD1 mutations in Portuguese patients with amyotrophic lateral sclerosis. *Neurobiol. Aging* **70**, 325.e7–325.e15 (2018).
102. Goldstein, O. *et al.* High frequency of C9orf72 hexanucleotide repeat expansion in amyotrophic lateral sclerosis patients from two founder populations sharing the same risk haplotype. *Neurobiol. Aging* **64**, 160.e1–160.e7 (2018).
103. Ogaki, K. *et al.* Analysis of C9orf72 repeat expansion in 563 Japanese patients with amyotrophic lateral sclerosis. *Neurobiol. Aging* **33**, 2527.e11–6 (2012).
104. Jang, J.-H. *et al.* Analysis of the C9orf72 hexanucleotide repeat expansion in Korean patients with familial and sporadic amyotrophic lateral sclerosis. *Neurobiol. Aging* **34**, 1311.e7–9 (2013).
105. Renton, A. E., Chiò, A. & Traynor, B. J. State of play in amyotrophic lateral sclerosis genetics. *Nat. Neurosci.* **17**, 17 (2013).

106. U.S. Food and Drug Administration. Coumadin product label 2017. (2017).
107. Gong, I. Y. *et al.* Determination of clinically therapeutic endoxifen concentrations based on efficacy from human MCF7 breast cancer xenografts. *Breast Cancer Res. Treat.* **139**, 61–69 (2013).
108. Drugs.com. Warfarin - FDA Prescribing Information. <https://www.drugs.com/pro/warfarin.html> (2018).
109. Mayo Clinic. Prothrombin time test Overview What you can expect. <https://www.mayoclinic.org/tests-procedures/prothrombin-time/about/pac-20384661> (2018).
110. American College of Cardiology. Warfarin (Coumadin ® , Jantoven ®) Considerations for Use *. <https://www.acc.org/tools-and-practice-support/clinical-toolkits/atrial-fibrillation-afib/anticoagulant-dosing-table/warfarin> (2019).
111. Pirmohamed, M. *et al.* A Randomized Trial of Genotype-Guided Dosing of Warfarin. *N. Engl. J. Med.* **369**, 2294–2303 (2013).
112. Suttie, J. W. & Jackson, C. M. Prothrombin structure, activation, and biosynthesis. *Physiol. Rev.* **57**, 1–70 (1977).
113. Whitlon, D. S., Sadowski, J. A. & Suttie, J. W. Mechanism of coumarin action: significance of vitamin K epoxide reductase inhibition. *Biochemistry* **17**, 1371–1377 (1978).
114. National Institutes of Health Clinical Centre Drug-Nutrient Interaction Task Force. Important Drug and Food Information Warfarin and Vitamin K.

http://www.cc.nih.gov/ccc/patient_education/drug_nutrient/coumadin1.pdf

(2012).

115. Wells, P. S., Holbrook, A. M., Crowther, N. R. & Hirsh, J. Interactions of Warfarin with Drugs and Food. *Ann. Intern. Med.* **121**, 676–683 (1994).
116. Lubetsky, A., Dekel-Stern, E., Chetrit, A., Lubin, F. & Halkin, H. Vitamin K intake and sensitivity to warfarin in patients consuming regular diets. *Thromb. Haemost.* **81**, 396–399 (1999).
117. Shen, G. *et al.* Warfarin traps human vitamin K epoxide reductase in an intermediate state during electron transfer. *Nat. Struct. Amp Mol. Biol.* **24**, 69 (2016).
118. Breckenridge, A., Orme, M., Wesseling, H., Lewis, R. J. & Gibbons, R. Pharmacokinetics and pharmacodynamics of the enantiomers of warfarin in man. *Clin. Pharmacol. Ther.* **15**, 424–430 (1974).
119. Fasco, M. J. & Principe, L. M. R- and S-Warfarin inhibition of vitamin K and vitamin K 2,3-epoxide reductase activities in the rat. *J. Biol. Chem.* **257**, 4894–4901 (1982).
120. Choonara, I. A., Haynes, B. P., Cholerton, S., Breckenridge, A. M. & Park, B. K. Enantiomers of warfarin and vitamin K1 metabolism. *Br. J. Clin. Pharmacol.* **22**, 729–732 (1986).
121. Zhang, Z., Fasco, M. J., Huang, Z., Guengerich, F. P. & Kaminsky, L. S. Human cytochromes P4501A1 and P4501A2: R-warfarin metabolism as a probe. *Drug Metab. Dispos.* **23**, 1339–1346 (1995).

122. Rettie, A. E. *et al.* Hydroxylation of warfarin by human cDNA-expressed cytochrome P-450: a role for P-450C9 in the etiology of (S)-warfarin-drug interactions. *Chem. Res. Toxicol.* **5**, 54–59 (1992).
123. Kaminsky, L. S. & Zhang, Z. Y. Human P450 metabolism of warfarin. *Pharmacol. Ther.* **73**, 67–74 (1997).
124. Higashi, M. K. *et al.* Association between CYP2C9 genetic variants and anticoagulation-related outcomes during warfarin therapy. *J. Am. Med. Assoc.* **287**, 1690–1698 (2002).
125. Joffe, H. V. *et al.* Warfarin dosing and cytochrome P450 2C9 polymorphisms. *Thromb. Haemost.* **91**, 1123–1128 (2004).
126. Rost, S. *et al.* Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. *Nature* **427**, 537–541 (2004).
127. Li, T. *et al.* Identification of the gene for vitamin K epoxide reductase. *Nature* **427**, 541–544 (2004).
128. Rieder, M. J. *et al.* Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *N. Engl. J. Med.* **352**, 2285–2293 (2005).
129. Cooper, G. M. *et al.* A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood* **112**, 1022–1027 (2008).
130. Takeuchi, F. *et al.* A Genome-Wide Association Study Confirms VKORC1, CYP2C9, and CYP4F2 as Principal Genetic Determinants of Warfarin Dose. *PLOS Genet.* **5**, e1000433 (2009).

131. Jorgensen, A. L., FitzGerald, R. J., Oyee, J., Pirmohamed, M. & Williamson, P. R. Influence of CYP2C9 and VKORC1 on Patient Response to Warfarin: A Systematic Review and Meta-Analysis. *PLoS ONE* **7**, (2012).
132. The International Warfarin Pharmacogenetics Consortium. Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data. *N. Engl. J. Med.* **360**, 753–764 (2009).
133. Kimmel, S. E. *et al.* A Pharmacogenetic versus a Clinical Algorithm for Warfarin Dosing. *N. Engl. J. Med.* **369**, 2283–2293 (2013).
134. PharmGKB. CYP2D6 Frequency Table.
https://api.pharmgkb.org/v1/download/file/attachment/CYP2D6_frequency_table.xlsx (2020).
135. PharmVar CYP2D6. *PharmVar CYP2D6*
<https://www.pharmvar.org/gene/CYP2D6>.
136. Nofziger, C. *et al.* PharmVar GeneFocus: CYP2D6. *Clin. Pharmacol. Ther.* **107**, 154–170 (2020).
137. Gaedigk, A. *et al.* The CYP2D6 activity score: translating genotype information into a qualitative measure of phenotype. *Clin. Pharmacol. Ther.* **83**, 234–242 (2008).
138. PharmGKB. CYP2D6 Diplotype-Phenotype Table.
https://api.pharmgkb.org/v1/download/file/attachment/CYP2D6_Diplotype_Phenotype_Table.xlsx (2020).
139. Zanger, U. M. *et al.* Comprehensive analysis of the genetic factors determining expression and function of hepatic CYP2D6. *Pharmacogenetics* **11**, 573–585 (2001).

140. Schroth, W. *et al.* Improved Prediction of Endoxifen Metabolism by CYP2D6 Genotype in Breast Cancer Patients Treated with Tamoxifen. *Front. Pharmacol.* **8**, 582–582 (2017).
141. Fleeman, N. *et al.* The clinical effectiveness and cost-effectiveness of genotyping for CYP2D6 for the management of women with breast cancer treated with tamoxifen: a systematic review. *Health Technol. Assess.* **15**, (2011).
142. Caudle, K. E. *et al.* Standardizing CYP2D6 Genotype to Phenotype Translation: Consensus Recommendations from the Clinical Pharmacogenetics Implementation Consortium and Dutch Pharmacogenetics Working Group. *Clin. Transl. Sci.* **13**, 116–124 (2020).
143. Sangar, M. C., Anandatheerthavarada, H. K., Martin, M. V., Guengerich, F. P. & Avadhani, N. G. Identification of genetic variants of human cytochrome P450 2D6 with impaired mitochondrial targeting. *Mol. Genet. Metab.* **99**, 90–97 (2010).
144. Sistonen, J. *et al.* CYP2D6 worldwide genetic variation shows high frequency of altered activity variants and no continental structure. *Pharmacogenet. Genomics* **17**, 93–101 (2007).
145. CYP2D6 Testing and Genotyping – Indications | Genelex. <https://www.genelex.com/test-menu/cyp2d6/>.
146. Sutiman, N. *et al.* Pharmacogenetics of UGT1A4 , UGT2B7 and UGT2B15 and Their Influence on Tamoxifen Disposition in Asian Breast Cancer Patients. *Clin. Pharmacokinet.* **55**, 1239–1250 (2016).
147. PharmVar: CYP2B6. <https://www.pharmvar.org/gene/CYP2B6>.

148. Coller, J. K. *et al.* The influence of CYP2B6, CYP2C9 and CYP2D6 genotypes on the formation of the potent antioestrogen Z-4-hydroxy-tamoxifen in human liver. *Br. J. Clin. Pharmacol.* **54**, 157–167 (2002).
149. PharmVar: CYP2C19. <https://www.pharmvar.org/gene/CYP2C19>.
150. 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).
151. Tucker, A. N. *et al.* Polymorphisms in cytochrome P4503A5 (CYP3A5) may be associated with race and tumor characteristics, but not metabolism and side effects of tamoxifen in breast cancer patients. *Cancer Lett.* **217**, 61–72 (2005).
152. Kiyotani, K. *et al.* A genome-wide association study identifies locus at 10q22 associated with clinical outcomes of adjuvant tamoxifen therapy for breast cancer patients in Japanese. *Hum. Mol. Genet.* **21**, 1665–1672 (2012).
153. Altshuler, D., Donnelly, P. & Consortium, T. I. H. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
154. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227–234 (2008).
155. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).
156. Onishi, H. *et al.* A genome-wide association study identifies three novel genetic markers for response to tamoxifen: A prospective multicenter study. *PLOS ONE* **13**, e0201606 (2018).

157. 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).
158. Hennig, E. E. *et al.* Limited predictive value of achieving beneficial plasma (Z)-endoxifen threshold level by CYP2D6 genotyping in tamoxifen-treated Polish women with breast cancer. *BMC Cancer* **15**, 570 (2015).
159. Yang, X. *et al.* Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. *Genome Res.* **20**, 1020–1036 (2010).
160. Wang, A., Stout, C. D., Zhang, Q. & Johnson, E. F. Contributions of Ionic Interactions and Protein Dynamics to Cytochrome P450 2D6 (CYP2D6) Substrate and Inhibitor Binding. *J. Biol. Chem.* **290**, 5092–5104 (2015).
161. 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 J. Integr. Biol.* **17**, 353–367 (2013).
162. Illumina. Genome-Wide DNA Analysis BeadChips. (2010).
163. Illumina. Infinium OmniExpress-24 v1.2 BeadChip. *Data Sheet Genotyping* 1–2 (2014).
164. Illumina. An Introduction to Next-Generation Sequencing Technology. (2017) doi:Pub No. 770-2012-008.
165. Hui, L., DelMonte, T. & Ranade, K. Genotyping Using the TaqMan Assay. *Curr. Protoc. Hum. Genet.* **56**, (2008).
166. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2017).

167. Weir, B. S. *et al.* GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* **28**, 3329–3331 (2012).
168. Clayton, D. *snpStats: SnpMatrix and XSnpMatrix classes and methods.* (2018).
169. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
170. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
171. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906 (2007).
172. Band, G. & Marchini, J. QCTOOL. <https://www.well.ox.ac.uk/gav/qctool/#overview>.
173. Freeman, C. & Marchini, J. GTOOL. <https://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html>.
174. Galesloot, T. E., van Steen, K., Kiemeneij, L. A. L. M., Janss, L. L. & Vermeulen, S. H. A Comparison of Multivariate Genome-Wide Association Methods. *PLOS ONE* **9**, 1–8 (2014).
175. Band, G. & Marchini, J. BGEN File Format. https://www.well.ox.ac.uk/~gav/bgen_format/.
176. Howie, B. N. & Marchini, J. Impute2. https://mathgen.stats.ox.ac.uk/impute/impute_v2.html (2014).

177. Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F. & Marchini, J. Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* **93**, 687–696 (2013).
178. Browning, S. R. & Browning, B. L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
179. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
180. Wang, Y., Lin, G., Li, C. & Stothard, P. Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle. *Springer Sci. Rev.* **4**, 79–98 (2017).
181. Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
182. Landis, S. H., Murray, T., Bolden, S. & Wingo, P. A. Cancer statistics, 1999. *CA. Cancer J. Clin.* **49**, 8–31 (1999).
183. Cancer Research UK. Breast Cancer Incidence by Sex and UK Country. <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/incidence-invasive#heading-Zero> (2015).
184. 1000 Genomes Project Consortium, T. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

185. Anderson, C. a *et al.* Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–73 (2010).
186. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
187. Turner, S. *et al.* Quality control procedures for genome-wide association studies. *Curr. Protoc. Hum. Genet.* 1–18 (2011) doi:10.1002/0471142905.hg0119s68.
188. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
189. Chapman, N. H. & Thompson, E. A. A model for the length of tracts of identity by descent in finite random mating populations. *Theor. Popul. Biol.* **64**, 141–150 (2003).
190. Tabangin, M. E., Woo, J. G. & Martin, L. J. The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proc.* **3**, S41 (2009).
191. Johnson, J. L., Abecasis, G. R., Heights, W. & Arbor, A. GAS Power Calculator: web-based power calculator for genetic association studies Department of Biostatistics and Center for Statistical Genetics , University of Michigan , 1420. *bioRxiv* (2017) doi:10.1038/ng.2507.S.
192. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209–213 (2006).
193. RStudio Team. *RStudio: Integrated Development Environment for R.* (RStudio, Inc., 2016).

194. Begum, F., Ghosh, D., Tseng, G. C. & Feingold, E. Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res.* **40**, 3777–3784 (2012).
195. Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* **33**, 177–182 (2003).
196. Turner, S. D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv* (2014) doi:10.1101/005165.
197. Devlin, B. & Roeder, K. Genomic control for association studies. *Am J Hum Genet* **65**, A83–A83 (1999).
198. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype Imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
199. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, (2009).
200. Illumina. The Omni Family of Microarrays. 0–2 (2010).
201. Download : Software : Sequence Read Archive : NCBI/NLM/NIH. <https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software>.
202. Krueger, F. Babraham Bioinformatics - Trim Galore! https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
203. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

204. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).
205. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).
206. *Picard toolkit. Broad Institute, GitHub repository* (Broad Institute, 2019).
207. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
208. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
209. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
210. Power Calculations: Quantitative Traits - Genome Analysis Wiki. https://genome.sph.umich.edu/wiki/Power_Calculations:_Quantitative_Traits.
211. Interpret Regression Coefficient Estimates - {level-level, log-level, level-log & log-log regression} - Curtis Kephart. <http://www.cazaar.com/ta/econ113/interpreting-beta>.
212. CYP2D6 Allele Functionality Table. *PharmGKB* https://api.pharmgkb.org/v1/download/file/attachment/CYP2D6_allele_functionality_reference.xlsx.
213. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinforma. Oxf. Engl.* **31**, 3555–3557 (2015).

214. Wang, D., Papp, A. C. & Sun, X. Functional characterization of CYP2D6 enhancer polymorphisms. *Hum. Mol. Genet.* **24**, 1556–1562 (2015).
215. Sanchez-Spitman, A. B. *et al.* The effect of rs5758550 on CYP2D6*2 phenotype and formation of endoxifen in breast cancer patients using tamoxifen. *Pharmacogenomics* **18**, 1125–1132 (2017).
216. Rekdal, C., Sjøttem, E. & Johansen, T. The Nuclear Factor SPBP Contains Different Functional Domains and Stimulates the Activity of Various Transcriptional Activators. *J. Biol. Chem.* **275**, 40288–40300 (2000).
217. Lyngsø, C. *et al.* Interaction between the Transcription Factor SPBP and the Positive Cofactor RNF4 AN INTERPLAY BETWEEN PROTEIN BINDING ZINC FINGERS. *J. Biol. Chem.* **275**, 26144–26149 (2000).
218. Elvenes, J. *et al.* Pax6 Represses Androgen Receptor-Mediated Transactivation by Inhibiting Recruitment of the Coactivator SPBP. *PLOS ONE* **6**, e24659 (2011).
219. Gburcik, V., Bot, N., Maggiolini, M. & Picard, D. SPBP Is a Phosphoserine-Specific Repressor of Estrogen Receptor α . *Mol. Cell. Biol.* **25**, 3421–3430 (2005).
220. Qiao, W. *et al.* Long-Read Single Molecule Real-Time Full Gene Sequencing of Cytochrome P450-2D6. *Hum. Mutat.* **37**, 315–323 (2016).
221. Fukunaga, K. *et al.* Determination of novel CYP2D6 haplotype using the targeted sequencing followed by the long-read sequencing and the functional characterization in the Japanese population. *J. Hum. Genet.* **66**, 139–149 (2021).

222. Liao, Y. *et al.* Nanopore sequencing of the pharmacogene CYP2D6 allows simultaneous haplotyping and detection of duplications. *bioRxiv* 576280 (2019) doi:10.1101/576280.